Adaptive Methods for Risk Calibration

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)

eingereicht an der
Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
M. Sc Weining Wang
geboren am in

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Ulrich Kamecke

Gutachter:

a. Prof. Dr. Wolfgang Karl Härdle

b. Prof. Dr. Vladimir Spokoiny

Tag des Kolloquiums: 2nd August 2012

Acknowledgements

Completing a PhD degree together with my master is a four-year and fruitful journey. I learned a lot on doing research and presenting results from it as well as on how to communicate and cooperate with people in the scientific community. I must greatly acknowledge my coauthors, teachers, and colleagues, who shared a lot of their experiences.

My first gratitude goes to my principal advisor, Dr. Wolfgang Karl Härdle. His influence on me is majorly on scientific research motivation. He provided me with the statistics vision and artistic thinking. He is good at motivating students to work hard and offered various opportunities to attend scientific conferences and workshops. He also made me connected with many world renowned scholars, from whom I gained inspirations for research and for life. He has been a strong and supportive adviser to me throughout my graduate school, while I can still have great freedom to pursue independent work.

Special thanks goes to the other advisor Dr. Vladimir Spokoiny for his mathematical insights and advice on structuring and editing paper. His enthusiasm and endurance in developing original and elegant statistical theorems inspired me a lot. I want to also thank Dr. Ya'acov Ritov for his ideas and inspirations, and for him being a role model for me as a statistician.

I am also grateful to Dr. Ostap Okhrin, Dr. Yarema Okhrin, Dr. Martin Odening, Dr. Brenda López Cabrera, Dr. Ihtiyor Bobojonov, Dr. Jianqing Fan, Dr. Philippe Rigollet, Dr. Lixing Zhu, Dr. Cheng-Der Fuh for their shared wisdoms. Their guidance is very helpful and I owe them my heart-felt appreciation.

Members of LvB chair of stastistics, and WIAS also deserve my sincerest thanks, their friendship and assistance has meant more to me than I could ever express. Meanwhile, I could not complete my work without the financial support from the Deutsche Forschungsgemeinschaft via SFB 649 "Economic Risk", Humboldt-Universität zu Berlin.

My friends in US, Taiwan and other parts of the world were sources of laughter, joy, and support. I wish to thank my family, my father, mother, grandfather, aunt and specially to my husband, Wei Cui. Their love provided me with inspiration and was my driving force. I sincerely wish I could show them just how much I love and appreciate them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This article includes four chapters. The first chapter is entitled "Local Quantile Regression", and its summary: Quantile regression is a technique to estimate conditional quantile curves. It provides a comprehensive picture of a response contingent on explanatory variables. In a flexible modeling framework, a specific form of the conditional quantile curve is not a priori fixed. This motivates a local parametric rather than a global fixed model fitting approach. A nonparametric smoothing estimate of the conditional quantile curve requires to balance between local curvature and stochastic variability. In the first essay, we suggest a local model selection technique that provides an adaptive estimate of the conditional quantile regression curve at each design point. Theoretical results claim that the proposed adaptive procedure performs as good as an oracle which would minimize the local estimation risk for the problem at hand. We illustrate the performance of the procedure by an extensive simulation study and consider a couple of applications: to tail dependence analysis for the Hong Kong stock market and to analysis of the distributions of the risk factors of temperature dynamics.

The second chapter is entitled "Tie the straps: uniform bootstrap confidence interval for additive models". It considers a bootstrap "coupling" technique for nonparametric robust smoothers and quantile regression, and verify the bootstrap improvement. To cope with curse of dimensionality, a different "coupling" bootstrap technique is developed for additive models with either symmetric error distributions and further extension to the quantile regression framework. Our bootstrap method can be used in many situations like constructing confidence intervals and bands. We demonstrate the bootstrap improvement in simulations and in applications to firm expenditures and the interaction of economic sectors and the stock market.

The third chapter is about "Hidden Markov structures for dynamic copulae". It focused on the issue: how to understand the dynamics of a high dimensional non-normal dependency structure. A Multivariate Gaussian or mixed normal based time varying models are limited in capturing important types of data features such as heavy tails, asymmetry, and nonlinear dependencies. This chapter aims at tackling this problem by building up a hidden Markov model (HMM) for hierarchical Archimedean copulae (HAC). The HAC constitute a wide class of models for high dimensional dependencies, and HMM is a statistical technique for describing regime switching dynamics. HMM applied to HAC flexibly models high dimensional non-Gaussian time series.

In this chapter we apply the expectation maximization (EM) algorithm for parameter estimation. Consistency results for both parameters and HAC structures are established in an HMM framework. The model is calibrated to exchange rate data with a VaR application. This example is motivated by a local adaptive analysis that yields a time varying HAC model. We compare the forecasting performance

with other classical dynamic models. In another, second, application we model a rainfall process. This task is of particular theoretical and practical interest because of the specific structure and required untypical treatment of precipitation data.

The fourth chapter is on "Localising temperature risk". On the temperature derivative market, modeling temperature volatility is an important issue for pricing and hedging. In order to apply pricing tools of financial mathematics, one needs to isolate a Gaussian risk factor. A conventional model for temperature dynamics is a stochastic model with seasonality and inter temporal autocorrelation. Empirical work based on seasonality and autocorrelation correction reveals that the obtained residuals are heteroscedastic with a periodic pattern. The object of this research is to estimate this heteroscedastic function so that after scale normalisation a pure standardised Gaussian variable appears. Earlier work investigated this temperature risk in different locations and showed that neither parametric component functions nor a local linear smoother with constant smoothing parameter are flexible enough to generally describe the volatility process well. Therefore, we consider a local adaptive modeling approach to find at each time point, an optimal smoothing parameter to locally estimate the seasonality and volatility. Our approach provides a more flexible and accurate fitting procedure of localised temperature risk process by achieving excellent normal risk factors.

Chapter 2

Local quantile regression

2.1 Introduction

Quantile regression is gradually developing into a comprehensive approach for the statistical analysis of linear and nonlinear response models. Since the rigorous treatment of linear quantile regression by Koenker & Bassett (1978), richer models have been introduced into the literature, among them are nonparametric, semiparametric and additive approaches. Quantile regression or conditional quantile estimation is a crucial element of analysis in many quantitative problems. In financial risk management, the proper definition of quantile based Value at Risk impacts asset pricing, portfolio hedging and investment evaluation, Engle & Manganelli (2004), Cai & Wang (2008) and Fitzenberger & Wilke (2006). In labor market analysis of wage distributions, education effects and earning inequalities are analyzed via quantile regression. Other applications of conditional quantile studies include, for example, conditional data analysis of children growth and ecology, where it accounts for the unequal variations of response variables, see James, Hastie & Sugar (2010).

In applications, the predominantly used linear form of the calibrated models is mainly determined by practical and numerical reasonings. There are many efficient algorithms (like sparse linear algebra and interior point methods) available, Portnoy & Koenker (1989),
Portnoy & Koenker (1997), Koenker & Ferreira (1999), and Koenker (2005), etc. However, the assumption of a linear parametric structure can be too restrictive in many applications. This observation spawned a stream of literature on nonparametric modeling of quantile regression, Yu & Jones (1998), Fan, Hu & Truong (1994), etc. One line of thought concentrated on different smoothing techniques, e.g. splines, kernel smoothing, etc.; see Fan & Gijbels (1996). Another line of

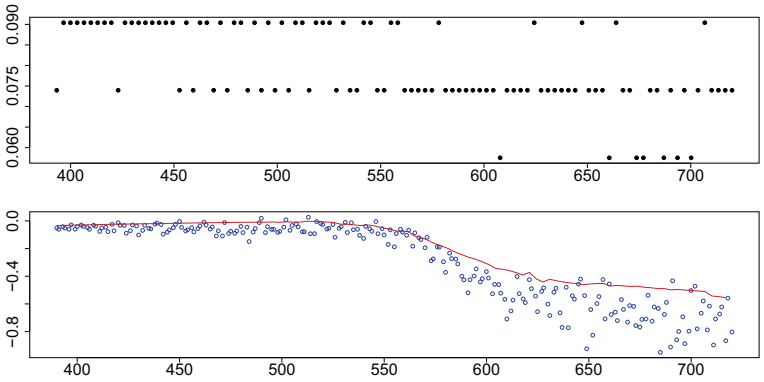


Figure 2.1: The bandwidth sequence (upper panel), plot of data and the estimated 90% quantile curve (lower panel)

literature considers structural semiparametric models to cope with curse of dimensionality, like, partial linear models, Härdle, Ritov & Song (2012), etc., additive models, Kong, Linton & Xia (2010), Horowitz & Lee (2005), etc; single index models, Wu, Yu & Yu (2010), Koenker (2010), etc. Yet another strand of literature has been involved in ultra-high dimensional situations where a careful variable selection technique needs to be implemented, Belloni & Chernozhukov (2010) and Koenker (2010). In most of the aforementioned papers on non and semiparametric quantile regression, a smoothing parameter selection is implicit, and it is mostly a consequence of theoretical assumptions like e.g. rates of convergence, but falls short in practical hints for real data applications. An important exception is the method for local nonparametric kernel smoothing by Yu & Jones (1998) and Cai & Xu (2008). They both propose a data driven choice of tuning parameter.

To address the limitations of the above mentioned literature on local model selection for nonparametric quantile regression, we aim at proposing with theoretical justification an adaptive local quantile regression algorithm that is easy to implement and works for a wide class of applications. The idea of this algorithm is to select tuning parameters locally by a sequence of likelihood ratio tests. The novelty lies in a local model selection technique with computable risk bounds. The main message is that the proposed algorithm is feasible and beneficial for quantile smoothing and helps in proposing alternatives to other models. As an example, consider Figure 2.1 which presents our results for analyzing the Lidar data set,

Ruppert, Wand & Carroll (2003). The presented quantile curve switches smoothness in the middle, and it is naturally reflected by the bandwidth sequence (upper panel) selected. In the presence of changing to sharper slope of the curve, the bandwidths get smaller to attain better approximations. This example shows that the paper algorithm can adaptively choose the bandwidth at each design point.

This article is organized as follows: In Section 2, we introduce the local model selection (LMS) procedure and lay down how to simulate critical values. In Section 3, Monte Carlo simulations are conducted to illustrate the proposed methodology. In Section 4, we apply our method on checking the tail dependency among portfolio stocks, and on estimation of quantile curves for temperature risk factors. In Section 5, we explain the main theorem on "Oracle" properties to support the validity of our tests, with the relevant assumptions, definitions and conditions in Appendix. In Section 6, we draw the conclusions. The technical details: 1, exponential risk bounds for conditional quantiles established using the representation of quantiles as Quasi Maximum Likelihood Estimation (qMLE)s of the asymmetric Laplace distribution, 2, theorems for the existence of critical values, 3, proof for "propagation", "stability" and "oracle" property are delegated to the Appendix.

2.2 Adaptive estimation procedure

This section introduces the considered problem and offers an adaptive estimation procedure.

2.2.1 Quantile regression model

Given the quantile level $\tau \in (0,1)$, the quantile regression model describes the following relation between the response Y and the regressor X:

$$I\!P(Y > f(x) \,|\, X = x) = \tau,$$

where $f(x)$ is the unknown *quantile regression function*. This function is the target of the analysis and it has to be estimated from independent observations $\{X_i, Y_i\}_{i=1}^n$. This relation can also be represented as

$$Y_i = f(X_i) + \varepsilon_i\,, \tag{2.1}$$

where the errors ε_i follow $I\!P(\varepsilon_i > 0 \,|\, X_i) = \tau$.

For simplicity of presentation, we consider a univariate regressor $X \in I\!R^1$ in this paper, an extension to the d-dimensional case $X \in I\!R^d$ with $d > 1$ is straightforward.

2.2.2 A qMLE View on Quantile Estimation

The quantile function $f(\cdot)$ in (2.1) is usually recovered by minimizing the sum

$$\sum_{i=1}^{n} \rho_\tau\{Y_i - f(X_i)\}, \tag{2.2}$$

over the class of all considered quantile functions $f(\cdot)$, where $\rho_\tau(u) \stackrel{\text{def}}{=} u\{\tau\, 1\!\!1(u \geq 0) - (1-\tau)\, 1\!\!1(u < 0)\}$. Such an approach is reasonable because the true quantile function $f(x)$ minimizes the expected value of the sum in (2.2). An important special case is given by $\tau = 1/2$. Then an estimate of $f(\cdot)$ is built as minimizer of the least absolute deviations (LAD) contrast $\sum |Y_i - f(X_i)|$.

The minimum contrast approach based on minimization of (2.2) can also be put in a quasi maximum likelihood framework. Assume that the residuals ε_i are (2.1) be i.i.d. and $\pi(x)$ is their negative log-density on $I\!R^1$. Then the joint log-density is given by the sum

$$-\sum \pi\big(Y_i - f(X_i)\big)$$

and its maximization is equivalent to minimization of the contrast (2.2) with a special density function coming from the *asymmetric Laplace distribution* (ALD):

$$\pi(u) = \pi_\tau(u) = \log\big\{\tau(1-\tau)\big\} - \rho_\tau(u), \quad -\infty < u < \infty. \tag{2.3}$$

The *parametric approach* (PA) additionally assumes that the quantile regression function $f(\cdot)$ belongs to a parametric family of functions $\big\{f_{\boldsymbol{\theta}}(x),\, \boldsymbol{\theta} \in \Theta\big\}$, where Θ is a subset of the p-dimensional Euclidean space. Equivalently,

$$f(x) = f_{\boldsymbol{\theta}^*}(x),$$

where $\boldsymbol{\theta}^*$ is the true parameter which is usually the target of estimation.

Examples are a constant model:

$$f_{\theta^*}(x) \equiv \theta_0,$$

with $\theta^* = \theta_0$ or a linear model:

$$f_{\boldsymbol{\theta}^*}(x) = \theta_0 + \theta_1 x,$$

with $\boldsymbol{\theta}^* = (\theta_0, \theta_1)^\top$.

Denote by $I\!P_{\boldsymbol{\theta}}$ the parametric measure on the observation space which corresponds to the regression model (2.1) with $f(\cdot) \equiv f_{\boldsymbol{\theta}}(\cdot)$ and with the i.i.d. errors

ε_i following the asymmetric Laplace distribution (2.3). Then the log-likelihood $L(\boldsymbol{\theta}) = L(\boldsymbol{Y}, \boldsymbol{\theta})$ for $I\!P_{\boldsymbol{\theta}}$ can be written as

$$L(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \log\bigl\{\tau(1-\tau)\bigr\} \sum_{i=1}^{n} 1 - \sum_{i=1}^{n} \rho_\tau\{Y_i - f_{\boldsymbol{\theta}}(X_i)\} \tag{2.4}$$

and the qMLE $\widetilde{\boldsymbol{\theta}}$ maximizes $L(\boldsymbol{\theta})$, or, equivalently minimizes the contrast $\sum_{i=1}^{n} \rho_\tau\{Y_i - f_{\boldsymbol{\theta}}(X_i)\}$ over all $\boldsymbol{\theta} \in \Theta$.

The described parametric construction is based on two assumptions: one is about the error distribution (2.3) and the other one is about the shape of the regression function f. However, it is only used for motivating our approach. Our theoretical study will be done under the true data distribution which follows mild regularity conditions. The next section explains how a smooth regression function f can be modeled by a flexible local parametric assumption.

2.2.3 Local polynomial qMLE

The *global PA* $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$ can be too restrictive in many applications. In what follows, we consider the local approach. Let a point x be fixed. The *local PA* at a point $x \in I\!R$ only requires that the quantile regression function $f(\cdot)$ can be approximated by a parametric function $f_{\boldsymbol{\theta}}(\cdot)$ from the given family in a vicinity of x. Below we fix a family of polynomial functions of degree p leading to the usual Taylor approximation:

$$f(u) \approx f_{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \theta_0 + \theta_1(u-x) + \ldots + \theta_p(u-x)^p/p! \tag{2.5}$$

for $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_p)^\top$. The corresponding parametric model can be written as

$$Y_i = \Psi_i^\top \boldsymbol{\theta} + \varepsilon_i\,, \tag{2.6}$$

where $\Psi_i = \{1, (X_i - x), (X_i - x)^2/2!, \ldots, (X_i - x)^p/p!\}^\top \in I\!R^{p+1}$.

A *local likelihood approach* at x is specified by a *localizing scheme* W given by a collection of weights w_i for $i = 1, \ldots, n$. The weights w_i vanish for points X_i lying outside a vicinity of the point x. A standard proposal for choosing the weights W is $w_i = K_{\text{loc}}\{(X_i - x)/h\}$, where $K_{\text{loc}}(\cdot)$ is a *kernel function* with a compact support, while h is a *bandwidth* controlling the degree of localization.

Define now the local log-likelihood at x by

$$L(W, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \log \tau(1-\tau) \sum_{i=1}^{n} w_i - \sum_{i=1}^{n} \rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}) w_i\,.$$

This expression is similar to the global log-likelihood in (2.4), but each summand in $L(W, \boldsymbol{\theta})$ is multiplied with the weight w_i, so only the points from the local vicinity of x contribute to $L(W, \boldsymbol{\theta})$. Note that this local log-likelihood depends on the central point x via the structure of the basis vectors Ψ_i and via the weights w_i. The corresponding local qMLE at x is defined via maximization of $L(W, \boldsymbol{\theta})$:

$$\begin{aligned} \widetilde{\boldsymbol{\theta}}(x) &= \{\widetilde{\theta}_0(x), \widetilde{\theta}_1(x), \ldots, \widetilde{\theta}_p(x)\}^\top \\ &\stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(W, \boldsymbol{\theta}) \\ &= \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1} \rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}) w_i \,. \end{aligned} \tag{2.7}$$

The first component $\widetilde{\theta}_0(x)$ provides an estimator of $f(x)$, while $\widetilde{\theta}_m(x)$ is an estimator of the derivative $f^{(m)}(x)$, $m = 1, \ldots, p$.

2.2.4 Selection of a Pointwise Bandwidth

The choice of bandwidth h is an important issue in implementing (2.7). One can reduce the variance of the estimation by increasing the bandwidth, but at a price of possibly inducing more modeling bias measured by the accuracy of approximation in (2.5); see Figure 2.2.

A desirable choice of a bandwidth at a fixed point would strike a balance between the variance and the bias depending on the local shape of $f(\cdot)$ in the vicinity of x. Many approaches have been proposed along this line; **see e.g. ???** However, their justification and implementation is based on some asymptotic arguments and require large samples. Here we propose a pointwise bandwidth selection technique based on finite sample theory.

Our basic setup of the algorithm is described as follows. First one fix a finite ordered set of possible bandwidths $h_1 < h_2 < \ldots < h_K$, where h_1 is very small, while h_K should be a global bandwidth of order of the design range. The bandwidth sequence can be taken geometrically increasing of the form $h_k = ab^k$ with fixed $a > 0$, $b > 1$, and $n^{-1} < ab^k < 1$ for $k = 1, \ldots, K$ (A.2.). The total number K of the candidate bandwidths is then at most logarithmic in the sample size n. This value enters in the oracle risk bound and the suggested choice ensures that the adaptive procedure is nearly efficient up to a log-factor in the estimation accuracy. Accordingly, the sequence of ordered weighting schemes $W^{(k)} = (w_1^{(k)}, w_2^{(k)}, \ldots, w_n^{(k)})^\top$ is defined via $w_i^{(k)} \stackrel{\text{def}}{=} K_{\text{loc}}\{(x - X_i)/h_k\}$. This leads

to a family of estimates $\widetilde{\boldsymbol{\theta}}_1(x), \widetilde{\boldsymbol{\theta}}_2(x), \ldots, \widetilde{\boldsymbol{\theta}}_K(x)$ with

$$\widetilde{\boldsymbol{\theta}}_k(x) = \operatorname*{argmax}_{\boldsymbol{\theta}} L(W^{(k)}, \boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1} \rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}) w_i^{(k)} . \tag{2.8}$$

The proposed selection procedure is similar in spirit to Lepski, Mammen & Spokoiny (1997).

If the underlying quantile regression function is smooth, one can expect a good quality of approximation (2.5) for a large bandwidth values among $\{h_k\}_{k=1}^K$. Moreover, if the approximation is good for one bandwidth, it will be also suitable for all smaller bandwidths. So, if we observe a significant difference between the estimate $\widetilde{\boldsymbol{\theta}}_k(x)$ corresponding to the bandwidth h_k and an estimate $\widetilde{\boldsymbol{\theta}}_\ell(x)$ corresponding to a smaller bandwidth h_ℓ, this is an indication that the approximation (2.5) for the window size h_k becomes too rough. This justifies the following procedure. Start with the smallest bandwidth h_1. For any $k > 1$, compute the local qMLE $\widetilde{\boldsymbol{\theta}}_k(x)$ and check it whether it is consistent with all the previous estimates $\widetilde{\boldsymbol{\theta}}_\ell(x)$ for $\ell < k$. If the consistency check is negative, the procedure terminates and select the latest accepted estimate.

The most important ingredient of the method is the consistency check. We follow the suggestion from Polzehl & Spokoiny (2006) and apply the localized likelihood ratio type test. More precisely, the local MLE $\widetilde{\boldsymbol{\theta}}_\ell(x)$ maximizes the log-likelihood value $L\bigl(W^{(\ell)}, \boldsymbol{\theta}\bigr)$, and the maximal value given by $\sup_{\boldsymbol{\theta}} L\bigl(W^{(\ell)}, \boldsymbol{\theta}\bigr) = L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr)$ is compared with the particular log-likelihood value $L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_k(x)\bigr)$, where the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ is obtained by maximizing the other local log-likelihood function $L(W^{(k)}, \boldsymbol{\theta})$. The difference $L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) - L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_k(x)\bigr)$ is always non-negative. The check rejects $\widetilde{\boldsymbol{\theta}}_k(x)$ if this difference is too large, that is, if it exceeds any specified critical value for any $\ell < k$. Equivalently one can say that the test checks whether $\widetilde{\boldsymbol{\theta}}_k(x)$ belongs to the confidence sets $\mathcal{E}_\ell(\mathfrak{z})$ of $\widetilde{\boldsymbol{\theta}}_\ell(x)$:

$$\mathcal{E}_\ell(\mathfrak{z}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\theta} : L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) - L\bigl(W^{(\ell)}, \boldsymbol{\theta}\bigr) \le \mathfrak{z}\bigr\}.$$

A great advantage of the likelihood ratio test is that the critical value $\mathfrak{z}$ can be selected universally. This is justified by the Wilks phenomenon: the likelihood ratio test statistics is nearly χ^2 and its asymptotic distribution depends only on the dimension of the parameter space. Unfortunately, these arguments do not apply for finite samples and under possible model misspecification and we offer later another way of fixing the critical values $\mathfrak{z}$ which is based on the so called propagation condition. We also allow that the width of the confidence set $\mathcal{E}_\ell(\mathfrak{z})$ depends on the index ℓ, that is, $\mathfrak{z} = \mathfrak{z}_\ell$. Our adaptation algorithm can be summarized as follows: At each step k, an estimator $\widehat{\boldsymbol{\theta}}_k(x)$ is constructed based

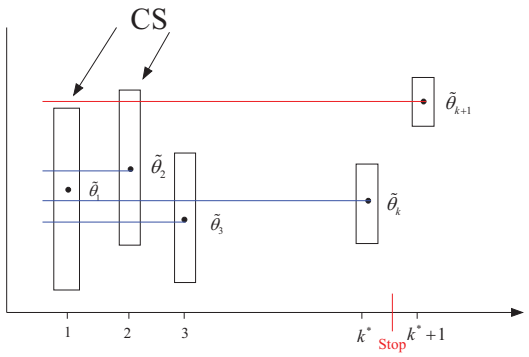


Figure 2.2: Demonstration of the local adaptive algorithm.

on the first k estimators $\widetilde{\boldsymbol{\theta}}_1(x), \ldots, \widetilde{\boldsymbol{\theta}}_k(x)$ by the following rule:

- Start with $\widehat{\boldsymbol{\theta}}_1(x) = \widetilde{\boldsymbol{\theta}}_1(x)$.
- For $k \geq 2$, $\widetilde{\boldsymbol{\theta}}_k(x)$ is accepted and $\widehat{\boldsymbol{\theta}}_k(x) \stackrel{\text{def}}{=} \widetilde{\boldsymbol{\theta}}_k(x)$, if $\widetilde{\boldsymbol{\theta}}_{k-1}(x)$ was accepted and

$$L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) - L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_k(x)\bigr) \leq \mathfrak{z}_\ell, \qquad \ell = 1, \ldots, k-1. \tag{2.9}$$

- The adaptive estimator $\widehat{\boldsymbol{\theta}}(x)$ is the latest accepted estimator after all K steps.

We also denote by $\widehat{\boldsymbol{\theta}}_k(x)$ is the latest accepted estimator after the first k steps. A visualization of the procedure is presented in Figure 2.2. The critical values $\mathfrak{z}_\ell$'s are selected by an algorithm based on the propagation condition explained in the next section.

2.2.5 Parameter Tuning by Propagation Condition

The practical implementation requires to fix the critical values of $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$. We apply the *propagation* approach which is an extension of the proposal from **?**. The idea of the approach is to tune the parameter of the procedure for one artificial parametric situation. Later we show that such defined critical value work well in the general setup and provide a nearly efficient estimation quality.

Similarly to Spokoiny (2009), the presented method can be viewed as a multiple testing procedure. This suggests to fix the critical values as in the general testing

theory by ensuring a prescribed performance under the null hypothesis. In our case, the null hypothesis corresponds to the pure parametric situation with $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$ in the equation (2.1). Moreover, we fix some particular distribution of the errors ε_i, our specific choice is the asymmetric Laplace distribution with the quantile parameter τ. Below in this section we denote by $I\!P_{\boldsymbol{\theta}^*}$ the data distribution under these assumptions.

For this artificial data generating process, all the estimates $\widetilde{\boldsymbol{\theta}}_k(x)$ should be consistent to each other and the procedure should not terminate at any intermediate step $k < K$. We call this effect as *propagation*: in the parametric situation, the degree of locality will be successfully increased until it reaches the largest scale. The critical values are selected to ensure the desired *propagation condition* which effectively means a "no false alarm" property: the selected adaptive estimate coincides in the most of cases with the estimate $\widetilde{\boldsymbol{\theta}}_K(x)$ corresponding to the largest bandwidth. The event $\{\widetilde{\boldsymbol{\theta}}_k(x) \neq \widehat{\boldsymbol{\theta}}_k(x)\}$ for $k \leq K$ is associated with a false alarm and the corresponding loss can be measured by the difference

$$L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr) \stackrel{\text{def}}{=} L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x)\bigr) - L\bigl(W^{(k)}, \widehat{\boldsymbol{\theta}}_k(x)\bigr).$$

The *propagation condition* postulates that the risk induced by such false alarms is smaller than the upper bound for the risk of the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ in the pure parametric situation:

$$I\!E_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr) \leq \alpha \mathcal{R}_r \qquad k = 2, \ldots, K, \tag{2.10}$$

where it holds for all $k \leq K$ and

$$I\!E_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr) \leq \mathcal{R}_r$$

Here α and r are two hyper-parameters. The role of α is similar to the significance level of a test, while r denotes the power of the loss function. It is worth mentioning that

$$I\!E_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr) \to I\!P_{\boldsymbol{\theta}^*}\{\widetilde{\boldsymbol{\theta}}_k(x) \neq \widehat{\boldsymbol{\theta}}_k(x)\}, \qquad r \to 0.$$

The critical values $\{\mathfrak{z}_k\}_{k=1}^{K-1}$ enter implicitly in the propagation condition: if the false alarm event $\{\widetilde{\boldsymbol{\theta}}_k(x) \neq \widehat{\boldsymbol{\theta}}_k(x)\}$ happens too often, it suggests that some of the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$ are too small. Note that (2.10) relies on the artificial parametric model $I\!P_{\boldsymbol{\theta}^*}$ instead of the true model $I\!P$. The point $\boldsymbol{\theta}^*$ here can be selected arbitrarily, e.g. $\boldsymbol{\theta}^* = 0$. This fact relies on linear parametric structure of the model (2.6) and is justified by the following simple lemma.

Lemma 1. *The distribution of* $L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr)$ *and of* $L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr)$ *under* $I\!P_{\boldsymbol{\theta}^*}$ *does not depend on* $\boldsymbol{\theta}^*$.

Proof. Under PA $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$, it holds $Y_i - f(X_i) = Y_i - \Psi_i^\top \boldsymbol{\theta}^* = \varepsilon_i$ and hence,

$$L(W^{(k)}, \boldsymbol{\theta}) = \log\{\tau(1-\tau)\} \sum_{i=1}^n w_i^{(k)} + \sum_{i=1}^n \rho_\tau\big(\varepsilon_i - \Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\big) w_i^{(k)}.$$

A simple inspection of this formula yields that the distribution of $L(W^{(k)}, \boldsymbol{\theta})$ only depends on $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. In other words, we can use the free parameter $\boldsymbol{u} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ whatever $\boldsymbol{\theta}^*$ is, e.g. $\boldsymbol{\theta}^* \equiv 0$. The same argument applies to the difference $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\big)$ for $\ell < k$. Moreover, $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\big)$ is a function of $\big\{L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\big)\big\}_{\ell=1}^k$, so the distribution of $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\big)$ does not depend on $\boldsymbol{\theta}^*$. □

A choice of critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ can be implemented in the following way:

- Consider first only $\mathfrak{z}_1$ and fix $\mathfrak{z}_2 = \ldots = \mathfrak{z}_{K-1} = \infty$, leading to the estimates $\widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1, x)$ for $k = 2, \ldots, K$. The value $\mathfrak{z}_1$ is selected as the minimal one for which

$$\frac{1}{\mathcal{R}_r} I\!\!E_{\boldsymbol{\theta}^*} L^r\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1, x)\big) \le \frac{\alpha}{K-1}, \qquad k = 2, \ldots, K. \quad (2.11)$$

- With selected $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$, set $\mathfrak{z}_{k+1} = \ldots = \mathfrak{z}_{K-1} = \infty$. Any particular value of $\mathfrak{z}_k$ would lead to the set of parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_k, \infty, \ldots, \infty$ and the family of estimates $\widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \ldots, \mathfrak{z}_k, x)$ for $m = k+1, \ldots, K$. Select the smallest $\mathfrak{z}_k$ ensuring

$$\frac{1}{\mathcal{R}_r} I\!\!E_{\boldsymbol{\theta}^*} L^r\big(W^{(m)}, \widetilde{\boldsymbol{\theta}}_m(x), \widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_k, x)\big) \le \frac{k\alpha}{K-1} \quad (2.12)$$

  for all $m = k+1, \ldots, K$.

Few remarks to the proposed algorithm.

a. A value $\mathfrak{z}_1$ ensuring (2.11) always exists because the choice $\mathfrak{z}_1 = \infty$ yields $\widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1, x) = \widetilde{\boldsymbol{\theta}}_k(x)$ for all $k \ge 2$.

b. The value $L^r\big(W^{(m)}, \widetilde{\boldsymbol{\theta}}_m(x), \widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_k, x)\big)$ from (2.12) only accumulates the losses associated with the false alarms at the first k steps of the procedure, since the other checks at further steps are always accepted because the corresponding critical values $\mathfrak{z}_{k+1}, \ldots \mathfrak{z}_{K-1}$ are set to infinity.

c. The accumulated risk bound $\frac{k\alpha}{K-1}$ grows at each step by $\alpha/(K-1)$. This value can be seen as maximal risk accepted at step k.

d. The value $\mathfrak{z}_k$ ensuring (2.12) always exists, because the choice $\mathfrak{z}_k = \infty$ yields

$$\widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_k, x) = \widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \mathfrak{z}_2, \ldots, \mathfrak{z}_{k-1}, x)$$

for all $m \geq k$.

e. All the computed values depend on the considered linear parametric model, the sequence bandwidths h_k and the quantile level τ. They also depend on the local point x via the basis vectors Ψ_i. However, under usual regularity conditions on the design $X_1, \ldots, X_n$, the dependency on x is rather minor. Therefore, the adaptive estimation procedure can be repeated at different points without reiterating the step of selecting the critical values.

2.3 Simulations

First, we check the critical values at different quantile levels ($\tau = 0.05, 0.5, 0.75, 0.95$) and for different noise distributions, (a) Laplace, b) normal and c) student $t(3)$ distribution). Also, we study how does misidentification of noise distribution affects critical values.

Second, we compare the performance of our local bandwidth algorithm with two other bandwidth selection techniques. One proposal is from Yu & Jones (1998), in which they consider a rule of thumb bandwidth based on the assumption that the quantiles are parallel, and another comes from Cai & Xu (2008), where an approach based on a nonparametric version of the Akaike information criterion (AIC) is implemented.

2.3.1 Critical Values

Table 2.1 shows the critical values with several choices of α and r with $\tau = 0.75$, $m = 10000$ Monte Carlo samples, and an bandwidth sequence $(8, 14, 19, 25, 30, 36, 41, 52) * 0.001$ scaled for the interval $[0, 1]$. Critical values decrease when α increases, and increase when r increases, and the last 3 bandwidths equal to 0, which is natural, as by increasing the bandwidth, the variance of estimator decreases, and the size of the confidence set follows.

The bandwidth sequence in Table 2.2 displays critical values for different τ, with $\alpha = 0.25$, $r = 0.5$, $m = 10000$ Monte Carlo samples, a bandwidth sequence $\mathfrak{H}_1 = (8, 14, 19, 25, 30, 36, 41, 52) * 0.001$, and $\mathcal{N}(0, 1)$ noise. Critical values are roughly of the same level with respect to different τ.

Table 2.1: Critical Values with different r and α

$\alpha = 0.25,$	$r = 0.5$	6.123	2.333	0.987	3.678e-05	0.000
$\alpha = 0.5,$	$r = 0.5$	4.616	1.578	0.357	2.472e-05	0.000
$\alpha = 0.6,$	$r = 0.5$	3.203	0.679	0.025	0.006	7.278e-05
$\alpha = 0.25,$	$r = 0.75$	9.127	3.288	1.031	0.126	5.675e-05
$\alpha = 0.25,$	$r = 1$	12.75	4.280	1.224	1.095e-04	0.000

Table 2.2: Critical Values with Different τ

$\tau = 0.05$	6.464	2.204	0.620	3.345e-05	0.000
$\tau = 0.5$	7.997	3.089	0.986	0.300e-05	0.000
$\tau = 0.75$	9.203	3.910	1.106	0.123	7.254e-05
$\tau = 0.95$	8.589	5.452	1.904	0.334	1.203e-05

Table 2.3 displays the critical values for three alternative bandwidth sequences, i.e.
$\mathfrak{H}_2 = (8, 16, 25, 36, 49, 63, 79, 99) * 0.001$, $\mathfrak{H}_3 = (5, 8, 14, 19, 27, 36, 46, 58) * 0.001$ and
$\mathfrak{H}_1 = (8, 14, 19, 25, 30, 36, 41, 52) * 0.001$, with $\alpha = 0.25$, $r = 0.5$, and $\tau = 0.85$.
We see that critical values differ for different bandwidth sequences, α , r and τ , but they show the same patterns (finite and decreasing). This in fact guarantees that our algorithm works for difference choice of bandwidth sequences.

Table 2.3: Critical Values with Different Bandwidth Sequences

$\mathfrak{H}_1$	11.33	1.243	6.933e-05	0.000	0.000
$\mathfrak{H}_2$	18.39	6.479	2.230	0.469	8.738e-05
$\mathfrak{H}_3$	6.123	2.333	0.987	3.678e-05	0.000

We simulate from different data generating processes, namely the distribution of ε_i ($\pi(.)$) does not necessarily coincide with the likelihood (ALD_τ) taken to simulate critical values. Table 2.4 presents critical values simulated under $t(3)$, $\mathcal{N}(01)$ and ALD_τ . The critical values show the same trend with some differences, so we conclude that a misidentification of error distribution would not significantly contaminate the confidence sets.

In Table 2.5, critical values are shown in the same circumstances as in Table 2.4 for the local linear case. Since introducing one more variable (trend), critical values doubled or tripled compared to the local constant case. The behavior with respect

Table 2.4: Critical Values with Different Noise Distributions

$\mathcal{N}(0,1)$	11.50	4.924	2.514	1.313	2.765e-05
ALD_τ	14.05	6.554	3.304	1.443	5.879e-05
$t(3)$	15.42	8.707	2.370	0.342	3.898e-05

to tail functions stays the same.

Table 2.5: Critical Values with Different Noise Distributions in Local Linear Case

$\mathcal{N}(0,1)$	29.97	58.64	43.21	33.41	19.43	07.40
$ALD(0.5)$	45.28	74.51	66.43	50.42	31.42	13.50
$t(3)$	51.77	84.94	59.28	44.99	29.07	11.57

2.3.2 Comparison of Different Bandwidth Selection Techniques

We illustrate our proposal by considering $x \in [0,1]$, $\tau = 0.75$. The sample with ($n = 1000$) are simulated under three scenarios:

$$f^{[1]}(x) = \begin{cases} 0 & \text{if } x \in [0, 0.333]; \\ 8 & \text{if } x \in (0.333, 0666]; \\ -1 & \text{if } x \in (0.666, 1] \end{cases}$$

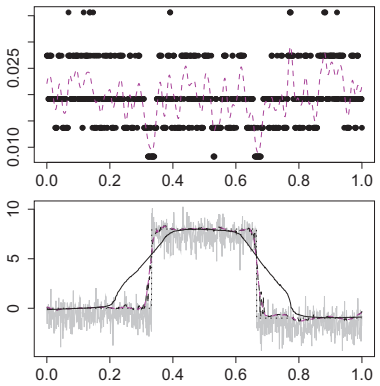
$$f^{[2]}(x) = 2x(1+x),$$

$$f^{[3]}(x) = \sin(k_1 x) + \cos(k_2 x)\, \mathbb{1}\{x \in (0.333, 0.666)\} + \sin(k_2 x)$$

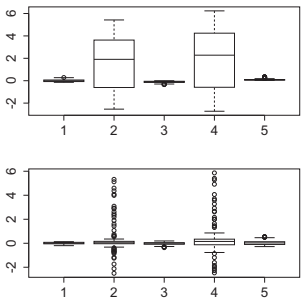
The noise distributions are: $\mathcal{N}(0, 0.03), ALD_\tau, t(3)$.

Figure 2.3 presents pictures on comparisons of different estimates in the local constant case. Figure 2.4 and 2.5 show in the local linear case the estimators of the functions ($\widehat{f(x)}$) and its first derivatives as well. Our technique provides closer fits to the true curve ($f(x)$) than methods with a global fixed bandwidth, especially in the presence of jump. Table 2.6, which shows the averaged absolute errors for the four methods, further confirms our conclusion.

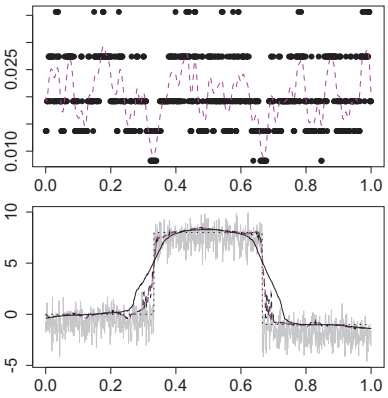
Table 2.7 offers further an analysis for misspecified error distributions. Specifically, to evaluate the accuracy of our estimation for error distributions generated differently than the ALD density. Table 2.7 gives L_1 errors between $\widehat{f}(\cdot)$ (with critical values simulated from ALD_τ) and $f(\cdot)$, from which we conclude that misspecification of error distributions would not contaminate our results significantly.



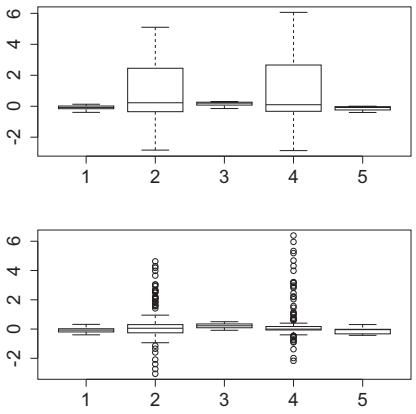
(a) Normal



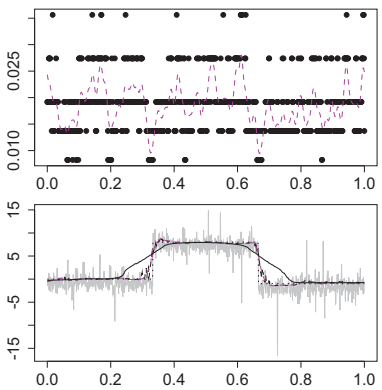
(b) Normal



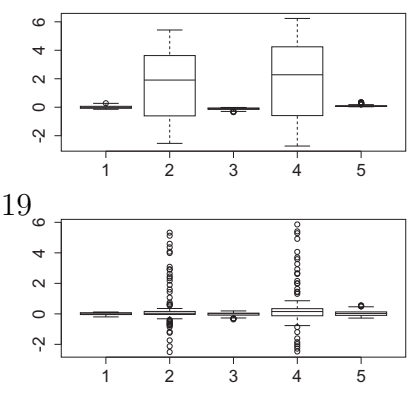
(c) ALD(0.5)



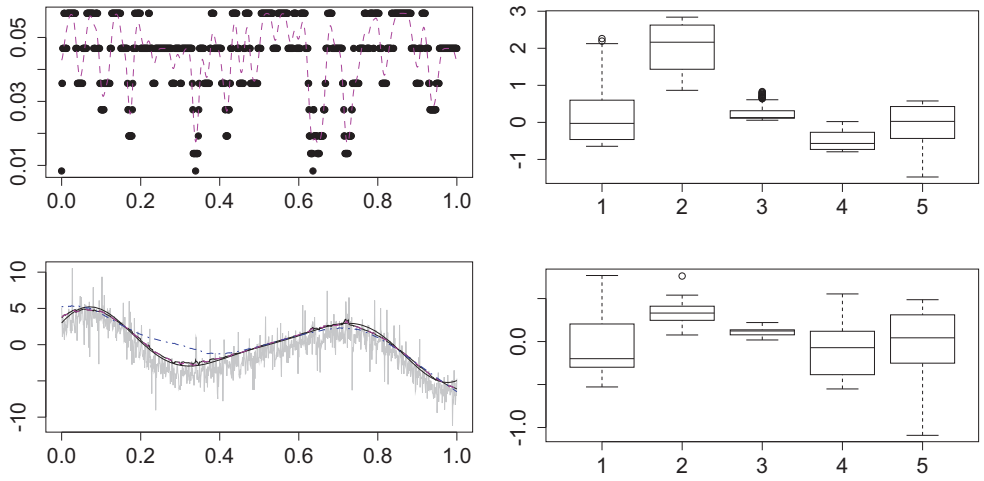
(d) ALD(0.5)



(e) t(3)

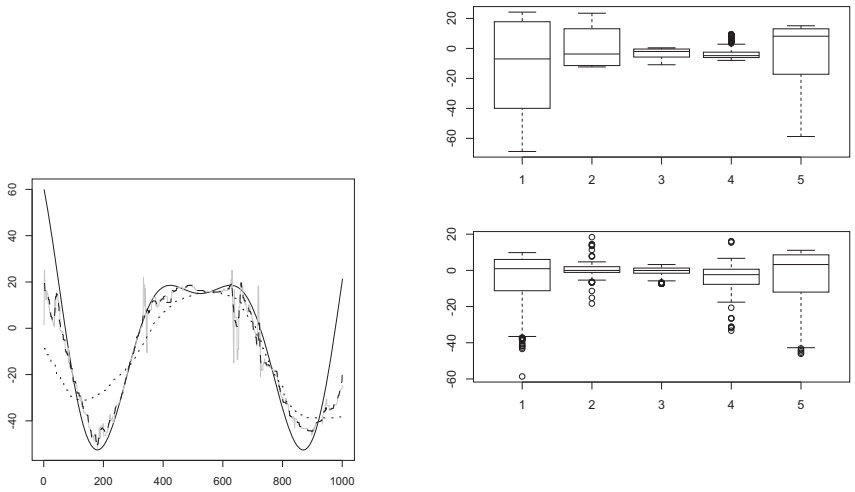


(f) t(3)



(a) (b)

Figure 2.4: The bandwidth sequence (upper left panel), the smoothed bandwidth sequence (dashed magenta); the observations (grey, lower left panel), the adaptive estimation of 0.75 quantile (dotted black), the true curve (solid black), the quantile smoother with fixed optimal bandwidth $= 0.063$ (dashed dotted blue), the estimation with adaptively smoothed bandwidth (dashed magenta); the blocked error of the adaptive estimator (lower right); the fixed estiamtor (upper right).



(a) (b)

Figure 2.5: The adaptive estimation of first derivative of the above quantile function (left panel grey), the true curve (solid black), the estimation with smoothed bandwidth (dashed black), the quantile smoother with fixed optimal bandwidth $= 0.045$ (dotted black); the blocked error of the adaptive estimator (lower right); the fixed estiamtor (upper right).

Table 2.6: Comparison of Monte Carlo errors, averaged over 1000 samples

	Fixed bandw	Local constant	Local linear	Fixed bandw (Cai)
$f^{[1]}(x)$	0.654	0.172	0.169	0.378
$f^{[2]}(x)$	0.206	0.008	0.008	0.245
$f^{[3]}(x)$	0.137	0.021	0.019	0.123

Table 2.7: Comparison of error mis-specification, errors are calculated averaged over 1000 samples

	Local constant $\{\mathcal{N}(0,1)\}$	Local constant $\{t(3)\}$	Local linear $\{\mathcal{N}(0,1)\}$
$f^{[1]}(x)$	0.252	0.220	0.169
$f^{[2]}(x)$	0.070	0.016	0.043
$f^{[3]}(x)$	0.009	0.021	0.019

2.4 Applications

In the study of financial products, it is very important to detect and understand tail dependence among underlyings such as stocks. In particular, the tail dependence structure represents the degree of dependence in the corner of the lower-left quadrant or upper-right quadrant of a bivariate distribution. Hauksson, Michel, Thomas, Ulrich & Gennady (2001) and Embrechts & Straumann (1999) provide a good access to the literature on tail dependence and Value at Risk. With the adaptive quantile technique, we provide an alternative approach to study tail dependence.

The correlation is calibrated from real data as given in Figure 2.6, where X is standardized return from stock "clpholdings" from Hong Kong Hangseng Index, and Y is return from stock "cheung kong". The conditional quantile function is linear, for example, $X_1 \in \mathcal{N}(u_1, \sigma_1)$ and $X_2 \in \mathcal{N}(u_2, \sigma_2)$, the conditional quantile function α is:

$$f(x) = \varphi^{-1}(\alpha)(\sigma_2 - \sigma_{12}^2/\sigma_1) + u_i + \sigma_{12}\sigma_2^{-1}(x - u_2).$$

Figure 2.6 and Figure 2.7 show the empirical conditional quantile curves actually deviate from the one calculated from normal distributions, which implies non normality. The motivation of adaptive bandwidth selection is clear to see from Figure 2.6 and Figure 2.7, the dependency structure change is more obvious compared with the fixed bandwidth curve. Moreover, the flexible adaptive curve is not likely to be a consequence of overfitting since it mostly lies in the confidence bands produced by fixed bandwidth estimation, see Härdle & Song (2010).

Figure 2.8 shows the first derivative curve for the above example. The curve gets more volatile while x increases until a drastically change, then it turns flat.

We measure the deviation from normality by accumulated L_1 distance to the normal fitting and examine different combination of stocks from Hong Kong Hangseng Index. The results is summarized in Table 2.8.

Table 2.8: Summary of deviation from normality

	Chalco	Cosco pacific	Bank of China
New world devo	0.252	0.220	0.169
Sino land	0.070	0.016	0.043
Swire pacific A	0.009	0.021	0.019

Another application of quantile function estimation is in temperature data analysis, which is of key interest for pricing temperature derivatives. Quantile regression can provide a more flexible and comprehensive approach to understand the temperature risk drivers defined in (5.6).

Denote daily temperature as $T \mapsto (t, j)$, with $t = 1, \cdots, \tau = 365$ days, $j = 0, \cdots, J$ years. The time series decomposition for $T_{t,j}$ is given as:

$$\begin{aligned} X_{t,j} &= T_{t,j} - \Lambda_t \\ X_{t,j} &= \sum_{l=1}^{L} \beta_l X_{t-l,j} + \sigma_t \eta_{t,j} \\ \eta_{t,j} &\sim \mathcal{N}(0,1), \\ \varepsilon_{t,j} &\stackrel{\text{def}}{=} \sigma_t \varepsilon_{t,j} \\ \widehat{\varepsilon}_{t,j} &\stackrel{\text{def}}{=} X_{365j+t} - \sum_{l=1}^{L} \widehat{\beta}_l X_{365j+t-l} \end{aligned} \tag{2.13}$$

where $T_{t,j}$ is the temperature at day t in year j, Λ_t denotes the seasonality effect and σ_t the seasonal volatility.

We are interested specifically in the stochastic risk drivers $\varepsilon_{t,j}$, Figure 2.9 presents a time series plot of $\widehat{\varepsilon}_{t,j}/\widehat{\sigma}_t$, and the estimated 90% quantile function. By zooming in the curve, we observe a very interesting phenomena: an changing of trend of the standardized residual over years.

To further understand the risk factors, we analyze the quantile functions of $\widehat{\varepsilon}^2_{t,j}$ over 12 years, and average over 4 years for comparison, see Figure 2.10 and Figure 2.11. The differences between Berlin and Kaoshiung are easy to see, the variance

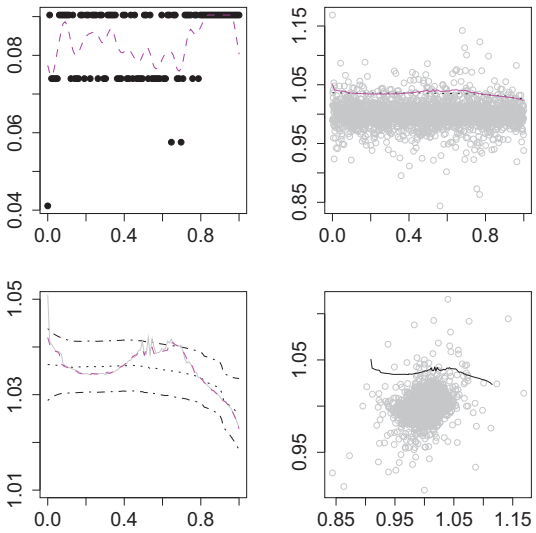


Figure 2.6: The bandwidth sequence with smoothed bandwidth curve(upper left panel), the smoothed bandwidth (dashed magenta); Scatter plot of stock returns (upper right panel), the adaptive estimation of 0.90 quantile (solid magenta), the quantile smoother with fixed optimal bandwidth $= 0.15$ (dotted black); fixed bandwidth curve (dotted black), adaptive bandwidth curve (grey), the estimation with smoothed bandwidth (dashed magenta), confidence band (dashed black) (lower left panel); adaptive bandwidth with normal scale (lower right).

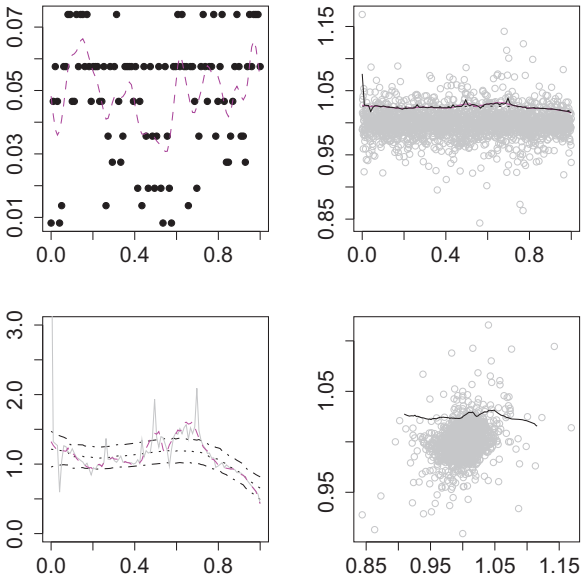


Figure 2.7: The bandwidth sequence with smoothed bandwidth curve (upper left panel); Scatter plot of stock returns (upper right panel), the adaptive estimation of 0.90 quantile (red), the quantile smoother with fixed optimal bandwidth = 0.19 (dotted black); fixed bandwidth curve (dotted black), adaptive bandwidth curve (grey), confidence bands (dotted dashed black) (lower left panel); adaptive bandwidth with normal scale (lower right panel)

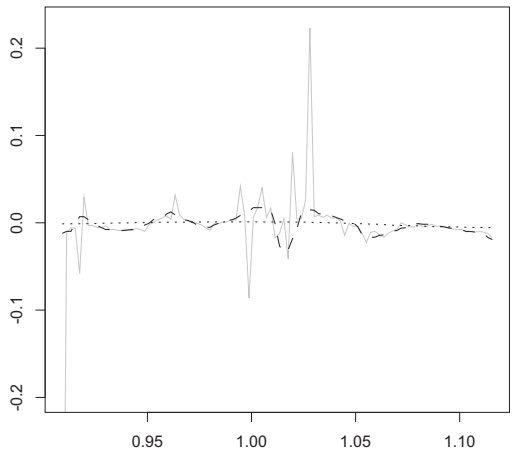


Figure 2.8: The adaptive trend curve (grey), smoothed adaptive curve (dashed black), estimation with fixed bandwidth (dotted black). $\tau = 0.90$

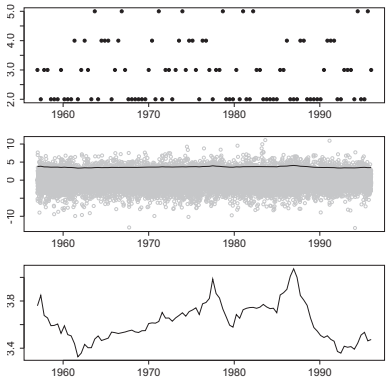


Figure 2.9: Plot of quantile curve for standardized weather residuals over 40 years at Berlin, 95% quantile, $1967 - 2006$. Selected bandwidths (upper), observations with estimated the quantile function (middle), the estimated the quantile function (lower).

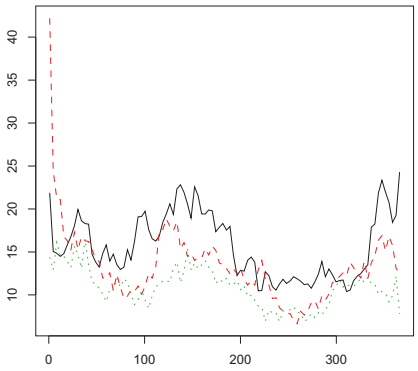


Figure 2.10: Estimated 90% quantile of variance functions, Berlin, average over $1995-1998$, $1999-2002$ (red), $2003-2006$ (green)

function has high value from Jan-Feb, while for Berlin the peaks come more in summer. Moreover, there is a tendency for Kaoshiung to be more volatile over time, but this phenomenon does not appear in Berlin.

In addition, our technique can also be used for estimating the function σ_t. We propose four methods: 1, Estimate the median curve of $\widehat{\varepsilon}_{t,j}$ using adaptive technique. 2, Take $\{\widehat{f}_{\varepsilon,0.75}-\widehat{f}_{\varepsilon,0.25}\}/1.34$ (1.34 is the inter quartile range of a standard normal distribution), where $\widehat{f}_{\varepsilon,0.75}$, $\widehat{f}_{\varepsilon,0.25}$ are the adaptive estimates. 3, Estimate the mean curve of $\widehat{\varepsilon}_{t,j}$ using adaptive bandwidth. 4, Estimate the mean function of $\widehat{\varepsilon}_{t,j}$ with fixed bandwidth. The aforementioned methods are compared by testing the normality of $\widehat{\eta}_{t,j} = \widehat{\varepsilon}_{t,j}/\widehat{\sigma}_t$. As according to our normal assumption on $\eta_{t,j}$, a good estimation for σ_t leads to normal standardized residuals $\widehat{\eta}_{t,j}$. Table 2.9 and 2.10 summarize statistics from the normality test of standardized residuals from three methods in Berlin and Kaoshiung. It can be seen that Berlin has more normal residuals than Kaoshiung. Method three is always better in getting more normal residuals, and method two is compatible with method three. It may be due to that quantiles at higher or lower levels are better to explain the extreme happened in volatility function. Method four performs not so well as it is with a fixed bandwidth. Therefore we conclude that our adaptive technique is useful in modeling temperature residuals.

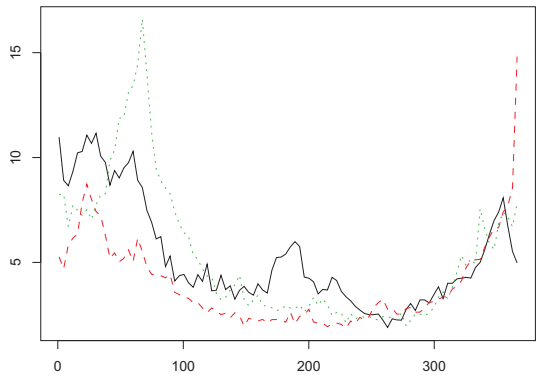


Figure 2.11: Estimated 90% quantile of variance functions, Kaoshiung, average over $1995-1998$, $1999-2002$ (red), $2003-2006$ (green)

Table 2.9: P-values of Normality Tests:Berlin

	AD	JB	KS
1	0.000	0.010	0.060
2	0.062	0.000	0.020
3	0.054	0.487	0.171
4	0.009	0.000	0.002

2.5 Finite Sample Theory

This section discusses some theoretical properties of the proposed estimate $\widehat{\boldsymbol{\theta}}(x) = \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)$. Here $\widehat{k} = \widehat{k}(x)$ is the index selected by the pointwise procedure from Section 2.2.4. The main "oracle" result shows that $\widehat{\boldsymbol{\theta}}(x)$ is *adaptive* in the sense that it provides nearly the same quality of estimation as the *oracle* estimate $\widetilde{\boldsymbol{\theta}}_{k^*}(x)$ which is the best in the family $\{\widetilde{\boldsymbol{\theta}}_k(x)\}_{k=1}^K$. A precise definition of k^* will be given below in term of the *modeling bias*.

2.5.1 Modeling Bias

This section explains the theoretical properties of the adaptive estimator $\widehat{\boldsymbol{\theta}}$ under a general data distribution. In such a case, a proper choice of a bandwidth becomes essential. The proposed approach for the bandwidth selection suggests to take larger and larger bandwidth until the linear parametric assumption is not significantly violated on the considered interval. The likelihood ratio test statistics

Table 2.10: P-values of Normality Tests:Kaoshiung

	AD	JB	KS
1	0.000	0.000	0.000
2	1.03e-05	0.077	0.043
3	2.37e-06	0.742	0.674
4	0.000	0.021	0.019

$L\bigl(W^{(\ell)}, \widetilde{\boldsymbol{\theta}}_\ell(x), \widetilde{\boldsymbol{\theta}}_k(x)\bigr)$ from (2.9) are used for this check. The formal definition of the best or oracle choice requires to introduce a measure for the deviation of the function $f(\cdot)$ from its best linear approximation $\Psi^\top \boldsymbol{\theta}$ on the interval of radius h_k considered at step k of the procedure. We follow Spokoiny (2009) which introduced the *modeling bias* for measuring the deviation from the linear parametric structure. Define P_i as the distribution of the observation Y_i. Let also $P_{i,s}$ be a shift of P_i by s, that is, the distribution of $Y_i - s$. Also denote $f_i = f(X_i)$ and $f_i(\boldsymbol{\theta}) = \Psi_i^\top \boldsymbol{\theta}$. In particular, P_{i,f_i} is the distribution of $\varepsilon_i \stackrel{\text{def}}{=} Y_i - f(X_i)$, so that its τ-quantile is zero. The underlying measure $\mathbb{P}$ is the product of the measures P_{i,f_i}. Under the linear PA $f(X_i) = f_{\boldsymbol{\theta}}(X_i)$, the corresponding measure $\mathbb{P}_{\boldsymbol{\theta}}$ is the product of the $P_{i,f_i(\boldsymbol{\theta})}$:

$$\mathbb{P} = \prod_{i=1}^n P_{i,f_i}\,, \qquad \mathbb{P}_{\boldsymbol{\theta}} = \prod_{i=1}^n P_{i,f_i(\boldsymbol{\theta})}\,.$$

The modeling bias at step k measures the deviation of the true quantile function f from the linear parametric one and it is defined as

$$\Delta_k \stackrel{\text{def}}{=} \inf_{\boldsymbol{\theta}} \Delta_k(\boldsymbol{\theta}),$$

$$\Delta_k(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^n \mathcal{K}\bigl(P_{i,f_i}, P_{i,f_i(\boldsymbol{\theta})}\bigr)\, 1\!\!1\{w_i^{(k)} > 0\}.$$

Here $\mathcal{K}(P,Q)$ is the Kullback-Leibler divergence between two measures P and Q. The quantity $\Delta_k(\boldsymbol{\theta})$ can be viewed as a kind of the Kullback-Leibler divergence between $\mathbb{P}$ and $\mathbb{P}_{\boldsymbol{\theta}}$ localized to the observations from the interval of radius h_k around x. The value Δ_k describes the quality of the best linear approximation on this interval. The *small modeling bias* (SMB) condition manifests that the value Δ_k does not exceed the prescribed quantity $\Delta > 0$, and the oracle choice of the bandwidth h_k is defined as the largest bandwidth among h_k for which the SMB

condition is satisfied:

$$k^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{k \le K} \{ \Delta_k \le \Delta \}. \tag{2.14}$$

Under the measure $I\!P_{\boldsymbol{\theta}^*}$, the estimate $\widetilde{\boldsymbol{\theta}}(x)$ is close to $\boldsymbol{\theta}^*$ in the sense that the confidence set $\mathcal{E}_k(\mathfrak{z}_{k^*})$ covers $\boldsymbol{\theta}^*$ with a high probability and the risk $I\!E_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \boldsymbol{\theta}^*\bigr)$ remains bounded by a fixed constant $\mathcal{R}_r$. The definition of the modeling bias based on the Kullback-Leibler divergence allows to translate this properties to the general case at cost of the additional factor e^{Δ}; see Lemma 2 below for more details. So, if Δ is small all the confidence or risk bounds continue to apply even in the local nonparametric situation.

2.5.2 "Oracle" Property

This section presents our main result called the oracle risk bound. The main message of this result is that the adaptive estimator $\widehat{\boldsymbol{\theta}}$ performs nearly as well as the best (oracle) estimator does. Let the bandwidth number k^* be defined by the SMB condition (2.14). Our first result describes the properties of the oracle estimator $\widetilde{\boldsymbol{\theta}}_{k^*}$. Under the parametric measure, its risk does not exceed $\mathcal{R}_r$. Under the SMB condition, its risk is of the same order up to a multiple e^{Δ}. The next result claims that the final estimator $\widehat{\boldsymbol{\theta}}$ is nearly as good as $\widetilde{\boldsymbol{\theta}}_{k^*}$ up to the factor $\mathfrak{z}_{k^*}\,\mathrm{e}^{\Delta}$.

Theorem 2.5.1. *Suppose* A.*1*–A.*5. Let* $\boldsymbol{\theta}$ *and* $k^* \le K$ *be such that* $\Delta_{k^*}(\boldsymbol{\theta}) \le \Delta$. *Then*

$$I\!E \log\biggl\{ 1 + \frac{L^r(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \boldsymbol{\theta})}{\mathcal{R}_r} \biggr\} \le \Delta + 1 \tag{2.15}$$

$$I\!E \log\biggl\{ 1 + \frac{L^r\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widehat{\boldsymbol{\theta}}(x)\bigr)}{\mathcal{R}_r} \biggr\} \le \alpha + \Delta + \log\bigl(1 + \frac{\mathfrak{z}_{k^*}}{\mathcal{R}_r}\bigr). \tag{2.16}$$

2.6 Conclusion

We propose an adaptive algorithm for nonparametric quantile estimation by local polynomial kernel regression with a flexible data-driven bandwidth selection. The procedure demonstrates a reasonable performance on the simulated and real data examples. The theory states the near optimality of the method for even for small or moderate samples in terms of the best or oracle estimator.

2.7 Appendix

The appendix collects the conditions, technical results, and the proofs. First we fix our assumptions. We assume independent observations $Y_1, \ldots, Y_n$. The results are stated for a deterministic design $X_1, \ldots, X_n$ under mild regularity conditions. The case of a random design can be considered by the usual conditioning argument. Given τ, the quantile function $f(\cdot)$ is defined by the relation $I\!P\{Y_i > f(X_i)\} = \tau$. To avoid ambiguous notation, we suppose that this equation has an unique solution for each i. The general case can be easily reduced to this one by standard arguments; see e.g. Koenker (2005). We also denote by P_i the distribution of the residual $\varepsilon_i = Y_i - f(X_i)$ and by $\pi_i(\cdot)$ its density. Below a point x is fixed and the target of estimation is the quantile $f(x)$. The local parametric approach requires to fix a localizing weighting scheme $W = (w_1, \ldots, w_n)$ and linear parametric family $f(\cdot, \boldsymbol{\theta})$ with $f(X_i, \boldsymbol{\theta}) = \Psi_i^\top \boldsymbol{\theta}$, where $\Psi_{i,m} = (X_i - x)^m / m!$ for $m = 0, 1, \ldots, p$.

Our theoretical study can be splitted into two parts. An essential and the most difficult part is done under the linear parametric assumption $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$. Then we extend the results to the case when this assumption is approximately fulfilled in a local vicinity of the central point x.

Below a family of localizing weighting schemes $W^{(k)} = \{w_i^{(k)}\}_{i=1}^n$ for $k = 1, \ldots, K$ is supposed to be fixed. Our standard proposal is $w_i^{(k)} = K_{\mathrm{loc}}\{(X_i - x)/h_k\}$ for a given kernel $K_{\mathrm{loc}}(\cdot)$ and a sequence of bandwidths $h_1 < h_2 < \ldots < h_K$. Define

$$D_k^2 \stackrel{\mathrm{def}}{=} -\frac{\partial^2 I\!E L(W^{(k)}, \boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} = \sum_{i=1}^n \Psi_i \Psi_i^\top \pi_i(0) w_i^{(k)} \tag{2.17}$$

$$V_k^2 \stackrel{\mathrm{def}}{=} \mathrm{Var}\{\nabla L(W^{(k)}, \boldsymbol{\theta}^*)\} = \tau(1-\tau) \sum_{i=1}^n \Psi_i \Psi_i^\top |w_i^{(k)}|^2, \tag{2.18}$$

$$N_k^{-1/2} \stackrel{\mathrm{def}}{=} \max_{i \le n} \sup_{\boldsymbol{\gamma} \in I\!R^p} \frac{|\boldsymbol{\gamma}^\top \Psi_i| w_i^{(k)}}{\|V_k \boldsymbol{\gamma}\|} \sqrt{\tau(1-\tau)},$$

the following conditions will be assumed for our results.

A.1 $\{Y_i\}_{i=1}^n$ are independent.

A.2 For some constants $0 < \mathfrak{u}_0 < \mathfrak{u} < 1$,

$$0 < \mathfrak{u}_0 \le \|D_k^{-1} D_{k-1}^2 D_k^{-1}\|_\infty \le \mathfrak{u} < 1.$$

A.3 For a constant $\mathfrak{a} > 0$ and all $k = 1, \ldots, K$, it holds

$$V_k^2 \leq \mathfrak{a}^2 D_k^2.$$

A.4 For some fixed $\delta < 1/2$ and $\rho > 0$,

$$\bigl|\pi_i(u)/\pi_i(0) - 1\bigr| \leq \delta, \qquad |u| \leq \rho.$$

A.5 The kernel function $K_{\mathrm{loc}}(\cdot)$ is supported on $[-1, 1]$, and is positive.

A.2 imposes a condition on choices of bandwidth sequence, it effectively requires that the bandwidth h_k grows geometrically with k. Condition A.3 is the local identifiability condition and it ensures that the local variability of the process $L(W^{(k)}, \boldsymbol{\theta})$ measured by the matrix V_k^2 is not significantly larger than the local information measured by the matrix D_k^2. A.4 only requires that the density functions $\pi_i(\cdot)$ are uniformly continuous in a vicinity of zero. In particular, the residuals can be unequally distributed. All the results below tacitly assume that the conditions A.1 – A.5 hold.

2.7.1 Uniform concentration of the MLEs $\widetilde{\boldsymbol{\theta}}_k(x)$

The first result explains the localization property of the estimates $\widetilde{\boldsymbol{\theta}}_k(x)$ from (2.8) under the linear parametric structure of the quantile function, that is, $f(X_i) = \Psi_i^\top \boldsymbol{\theta}^*$. With some value $\mathtt{r}_0$ fixed, define for each $k \leq K$ a local elliptic set

$$\Theta_k(\mathtt{r}_0) \stackrel{\mathrm{def}}{=} \bigl\{\boldsymbol{\theta} : \|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}_0\bigr\}$$

with $V_k = V(W^{(k)})$ from (2.18). The question under study is a proper choice of the radius $\mathtt{r}_0$ which ensures a prescribed small deviation probability for the event $\widetilde{\boldsymbol{\theta}}_k(x) \notin \Theta_k(\mathtt{r}_0)$ uniformly in $k \leq K$.

Below we use generic notation $C = C(\mathrm{A})$ to indicate that a constant C only depends on the constants from conditions A.1 – A.4 like $\mathfrak{a}$, ρ, δ, $\mathfrak{u}_0$, $\mathfrak{u}$, etc.

Theorem 2.7.1. *Suppose* $(E\mathtt{r})$ *and* $(\mathcal{L}\mathtt{r})$, *and there exist constants* $C_1 = C_1(\mathrm{A})$ *and* $C_2 = C_2(\mathrm{A})$ *such that the conditions*

$$\mathtt{r}_0^2 \geq C_1(\mathtt{x} + p), \qquad \rho^2 N_k \geq C_2(\mathtt{x} + p) \tag{2.19}$$

ensure for $k \leq K$

$$I\!P_{\boldsymbol{\theta}^*}\bigl\{\widetilde{\boldsymbol{\theta}}_k(x) \notin \Theta_k(\mathtt{r}_0)\bigr\} \leq 2\mathrm{e}^{-\mathtt{x}},$$

$$I\!E_{\boldsymbol{\theta}^*}\bigl[L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr)\, 1\!\!1\bigl\{\widetilde{\boldsymbol{\theta}}_k(x) \notin \Theta_k(\mathtt{r}_0)\bigr\}\bigr] \leq C(\mathrm{A})\mathrm{e}^{-\mathtt{x}}.$$

In particular, a choice $\mathtt{x} = \log(K) + \mathtt{x}_0$ and then $\mathtt{r}_0^2 \geq C_1(\mathtt{x} + p)$ ensures a dominating probability $1 - 2\mathrm{e}^{-\mathtt{x}_0}$ for the joint concentration event

$$\mathcal{A}_1 = \bigcap_{k=1}^{K} \bigl\{ \widetilde{\boldsymbol{\theta}}_k(x) \in \Theta_k(\mathtt{r}_0) \bigr\}.$$

In what follows we suppose that the values $\mathtt{x} = \log(K) + \mathtt{x}_0$ and $\mathtt{r}_0$ are fixed in a way that the probability of the set $\mathcal{A}_1$ is sufficiently close to 1. This allows to restrict ourselves to the case when each estimate $\widetilde{\boldsymbol{\theta}}_k(x)$ belongs to the local vicinity $\Theta_k(\mathtt{r}_0)$. The conditions in (2.19) require that $\mathtt{r}_0^2$ is of order $\log(K) + p$, and the local sample size N_k should be at least of the same order.

2.7.2 Uniform quadratic approximation of the local excess

The previous subsection stated that the chance for any of the estimator $\widetilde{\boldsymbol{\theta}}_k(x)$ lying outside the neighborhood $\Theta_k(\mathtt{r}_0)$ is small, therefore in this subsection, we focus on the stochastic behavior of $\widetilde{\boldsymbol{\theta}}_k$ in $\Theta_k(\mathtt{r}_0)$. The proposed estimation procedure is likelihood-based: all quantities are defined in terms of the quasi log-likelihood functions $L(W, \boldsymbol{\theta})$. Particularly, the properties of the *excess* $L\bigl(W, \widetilde{\boldsymbol{\theta}}(x), \boldsymbol{\theta}^*\bigr) \stackrel{\text{def}}{=} L\bigl(W, \widetilde{\boldsymbol{\theta}}(x)\bigr) - L(W, \boldsymbol{\theta}^*)$ plays a very important role in the whole method. The famous Wilks result claims that the excess is asymptotically χ^2_p. Unfortunately the local parametric approach for a narrow local neighborhoods of the point x leads to a relatively small effective sample size N, and the asymptotic results cannot be validated. In the contrary, the general parametric approach of Spokoiny (2011) allows to operate with finite samples and it can be directly applied to the local parametric situation.

It holds

$$\begin{aligned} \nabla L(W^{(k)}, \boldsymbol{\theta}^*) &= -\sum_{i=1}^{n} \rho'_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}^*) w_i^{(k)} \\ &= \sum_{i=1}^{n} \bigl\{ -\tau + \mathbb{1}\!\mathrm{I}(Y_i - \Psi_i^\top \boldsymbol{\theta}^* < 0) \bigr\} \Psi_i \, w_i^{(k)}. \end{aligned}$$

Further, for $\boldsymbol{\epsilon} = (\delta, \varrho)$ and $D_k^2 = D^2(W^{(k)})$ from (2.17), define

$$\begin{aligned} D^2_{\boldsymbol{\epsilon},k} &= D_k^2(1 - \delta) - \varrho V_k^2, \\ \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} &\stackrel{\text{def}}{=} D^{-1}_{\boldsymbol{\epsilon},k} \nabla L(W^{(k)}, \boldsymbol{\theta}^*), \end{aligned}$$

and similarly for $\underline{\boldsymbol{\epsilon}} \stackrel{\text{def}}{=} -\boldsymbol{\epsilon} = (-\delta, -\varrho)$. The values δ, ϱ are assumed to be small enough to ensure that $D^2_{\boldsymbol{\epsilon},k}$ is positive and the value

$$\alpha_{\boldsymbol{\epsilon},k} \stackrel{\text{def}}{=} \lambda_{\max}\bigl(I_p - D_{\boldsymbol{\epsilon},k} D^{-2}_{\underline{\boldsymbol{\epsilon}},k} D_{\boldsymbol{\epsilon},k}\bigr) \tag{2.20}$$

is small as well. Finally, define

$$\begin{aligned}\mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) &\stackrel{\text{def}}{=} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(W^{(k)}, \boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon},k}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2 \\ &= \boldsymbol{\xi}^\top_{\boldsymbol{\epsilon},k} D_{\boldsymbol{\epsilon},k}(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \|D_{\boldsymbol{\epsilon},k}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2/2\end{aligned}$$

and a similar definition for $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$.

Theorem 2.7.2. *It holds with conditions* (ED_0), (ED_1), $(\mathcal{L}_0)$,

$$\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) - \diamondsuit_{\boldsymbol{\epsilon},k} \le L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \le \mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) + \diamondsuit_{\boldsymbol{\epsilon},k}, \tag{2.21}$$

for all $\boldsymbol{\theta} \in \Theta_1(\mathtt{r}_0)$ *and all* $k \le K$. *Here* $\diamondsuit_{\boldsymbol{\epsilon},k}$ *are the random error terms which fulfill with some* $C_1(\mathrm{A})$ *and* $C_2(\mathrm{A})$ *the following conditions: for any* $\mathtt{x} > 0$ *with* $C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A}) \le \mathtt{y}_c$

$$\begin{aligned}I\!\!P_{\boldsymbol{\theta}^*}\bigl(\diamondsuit_{\boldsymbol{\epsilon},k}/\varrho > C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A})p\bigr) &\le C(\mathrm{A})\mathrm{e}^{-\mathtt{x}}, \\ I\!\!E_{\boldsymbol{\theta}^*}\bigl|\diamondsuit_{\boldsymbol{\epsilon},k}/\varrho\bigr|^r &\le C_r(\mathrm{A}),\end{aligned}$$

where $\mathtt{y}_c$ *is a constant of order* p.

The sandwiching result (2.21) for each k follows from Theorem 3.1 of Spokoiny (2011). It is only worth mentioning that the local sets $\Theta_k(\mathtt{r}_0)$ are embedded: $\Theta_1(\mathtt{r}_0) \supset \Theta_2(\mathtt{r}_0) \supset \ldots \supset \Theta_K(\mathtt{r}_0)$, so it suffices to check the bound (2.21) on $\Theta_1(\mathtt{r}_0)$ for each $k \le K$.

The majorization bound (2.21) yields that the the maximum of the process $L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ would also be sandwiched between the maximum of $\mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ up to a small random error term. Moreover, as $\mathbb{L}_{\boldsymbol{\epsilon}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ and $\mathbb{L}_{\underline{\boldsymbol{\epsilon}}}(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$ are quadratic, their maximum would be of a simple analytical form ($\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2$). The next result presents a bound on this squared norm.

Theorem 2.7.3. *There exist* $C_1(\mathrm{A})$ *and* $C_2(\mathrm{A})$ *such that for each* $\mathtt{x}$ *with* $C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A})p \le \mathtt{y}_c$ *and* $k \le K$, *it holds with conditions* (ED_0), (ED_1), $(\mathcal{L}_0)$,

$$I\!\!P_{\boldsymbol{\theta}^*}\bigl\{\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2 > C_1(\mathrm{A})\mathtt{x} + C_2(\mathrm{A})p\bigr\} \le 2\mathrm{e}^{-\mathtt{x}}.$$

Furthermore, for $r > 0$ *and* $k \le K$, *it holds*

$$I\!\!E\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^{2r} \le C_r(\mathrm{A}).$$

Similar to the result of Theorem 2.7.1, one can select a radius $\mathtt{r}_0$ such that the probability of the set $\mathcal{A}_2$ with

$$\mathcal{A}_2 = \bigcup_{k=1}^{K} \{ \| \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \| \leq \mathtt{r}_0 \}$$

sufficiently close to one. Below we restrict ourselves to the set $\mathcal{A}$ with $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2$.

The results of Theorem 2.7.2 and 2.7.3 have a number of important corollaries; cf. Spokoiny (2011).

Corollary 1. *It holds on* $\mathcal{A}$ *for every* $k \leq K$

$$\| \boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}},k} \|^2/2 - \diamondsuit_{\boldsymbol{\epsilon},k} \leq L(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*) \leq \| \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \|^2/2 + \diamondsuit_{\boldsymbol{\epsilon},k}. \tag{2.22}$$

Corollary 2. *It holds on* $\mathcal{A}$ *for every* $k \leq K$

$$\begin{aligned} \big\| D_{\boldsymbol{\epsilon},k} \big(\widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}^* \big) - \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \big\|^2 &\leq 4 \diamondsuit_{\boldsymbol{\epsilon},k} + \alpha_{\boldsymbol{\epsilon},k} \| \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \|^2, \\ \big\| D_{\boldsymbol{\epsilon},k} \big(\widetilde{\boldsymbol{\theta}}_k(x) - \boldsymbol{\theta}^* \big) \big\| &\leq 2 \diamondsuit_{\boldsymbol{\epsilon},k}^{1/2} + \big(1 + \alpha_{\boldsymbol{\epsilon},k}^{1/2} \big) \| \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \|. \end{aligned} \tag{2.23}$$

The result of Corollary 1 can be viewed as a non-asymptotic version of the Wilks Theorem. It claims that the twice excess $L(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*)$ can be approximated by the quadratic form $\| \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \|^2$. Moreover, the vector $\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}$ is asymptotically normal under usual assumptions by the central limit theorem, thus the twice excess is asymptotically χ^2_p. The next result describes some finite sample properties of $\| \boldsymbol{\xi}_{\boldsymbol{\epsilon},k} \|^2$.

One can summarize the obtained general results as follows. On the set $\mathcal{A}$ of dominating probability, each estimate $\widetilde{\boldsymbol{\theta}}_k(x)$ belongs to the local vicinity $\Theta_k(\mathtt{r}_0)$ which yields the bounds (2.22), (2.23). Moreover, the random quantities $\diamondsuit_{\boldsymbol{\epsilon},k}$ and $\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}$ obey the deviation and moment bounds of Theorem 2.7.2 and Theorem 2.7.3.

The conditions

Here we list the conditions from Spokoiny (2011) which are assumed to be fulfilled for each local likelihood $L(W^{(k)}, \boldsymbol{\theta})$, $k \leq K$. Some value $\mathtt{r}_0$ is assumed to be fixed for all conditions. It separates the local zone of local quadratic approximation and the large deviation zone. The assumption are stated under the true data distribution $I\!P$. However, we apply the assumptions only in the case of linear

parametric structure with $f(\cdot) \equiv f_{\boldsymbol{\theta}^*}(\cdot)$. Define

$$\begin{aligned}\zeta_k(\boldsymbol{\theta})) &\stackrel{\text{def}}{=} L(W^{(k)}, \boldsymbol{\theta}) - I\!\!E L(W^{(k)}, \boldsymbol{\theta}) \\ &= -\sum_{i=1}^{n} \Big\{ \rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta}) - I\!\!E\{\rho_\tau(Y_i - \Psi_i^\top \boldsymbol{\theta})\Big\} w_i^{(k)}.\end{aligned}$$

Also denote $\nabla \zeta_k(\boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}} \zeta(\boldsymbol{\theta})$. The following conditions are assumed to be fulfilled for each $k \le K$.

($\boldsymbol{ED_0}$) There exists a positive symmetric matrix V_k^2, and constants $\mathtt{g} > 0$ and $\nu_0 \ge 1$ such that $\mathrm{Var}\big\{\nabla \zeta_k(\boldsymbol{\theta})\big\} \le V_k^2$ and for all λ with $|\lambda| \le \mathtt{g}$,

$$\sup_{\boldsymbol{\gamma} \in I\!R^{p+1}} \log I\!\!E_{\boldsymbol{\theta}^*} \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla \zeta_k(\boldsymbol{\theta}^*)}{\|V_k \boldsymbol{\gamma}\|} \right\} \le \nu_0^2 \lambda^2 / 2,$$

With this matrix V_k, define the local set

$$\Theta_k(\mathtt{r}) = \{\boldsymbol{\theta} : \|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}$$

($\boldsymbol{ED_1}$) For each $\mathtt{r} \le \mathtt{r}_0$, there exists a constant $\varrho(\mathtt{r}) \le 1/2$ such that it holds for all $\boldsymbol{\theta} \in \Theta_k(\mathtt{r}_0)$ and $|\lambda| \le \mathtt{g}$:

$$\sup_{\boldsymbol{\gamma} \in I\!R^{p+1}} \log I\!\!E_{\boldsymbol{\theta}^*} \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \{\nabla \zeta_k(\boldsymbol{\theta}) - \nabla \zeta_k(\boldsymbol{\theta}^*)\}}{\varrho(\mathtt{r}) \|V_k \boldsymbol{\gamma}\|} \right\} \le \nu_0^2 \lambda^2 / 2,$$

($\boldsymbol{\mathcal{L}_0}$) *There are a positive matrix D_k and for each $\mathtt{r} \le \mathtt{r}_0$ and a constant $\delta(\mathtt{r}) \le 1/2$, such that it holds for all $\boldsymbol{\theta} \in \Theta_k$,*

$$\left| \frac{-2 I\!\!E L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} - 1 \right| \le \delta(\mathtt{r}),$$

($\boldsymbol{E\mathtt{r}}$) For any $\mathtt{r} \ge \mathtt{r}_0$, there exist a value $\mathtt{g}(\mathtt{r}) > 0$ and a constant ν_0 such that for all $\lambda \le \mathtt{g}(\mathtt{r})$,

$$\sup_{\boldsymbol{\gamma} \in I\!R^{p+1}} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \log I\!\!E_{\boldsymbol{\theta}^*} \exp\left\{ \lambda \frac{\boldsymbol{\gamma}^\top \nabla \zeta_k(\boldsymbol{\theta})}{\|V_k \boldsymbol{\gamma}\|} \right\} \le \nu_0^2 \lambda^2 / 2.$$

($\boldsymbol{\mathcal{L}\mathtt{r}}$) For each $\mathtt{r} \ge \mathtt{r}_0$ and any $\boldsymbol{\theta}$ with $\|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$,

$$\frac{-I\!\!E_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \ge \mathtt{b}(\mathtt{r}) > 0,$$

Conditions $(ED_0), (ED_1), (\mathcal{L}_0)$ are local conditions which should be applied on the local set $\Theta_k(\mathtt{r}_0)$, while $(\mathcal{L}\mathtt{r}), (E\mathtt{r})$ are global conditions which we apply on the complement of $\Theta_k(\mathtt{r}_0)$. Also $(ED_0), (ED_1), (E\mathtt{r})$ are smoothness or moment assumptions on the log likelihood process, and the conditions $(\mathcal{L}_0), (\mathcal{L}\mathtt{r})$ ensure the identifiability properties.

Proof of $(E\mathtt{r})$, (ED_0) and (ED_1).

First we check $(E\mathtt{r})$. It holds by definition

$$\begin{aligned}\nabla\zeta_k(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \Psi_i \big[\mathbb{1}\!\!1(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0) - I\!P(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0) \big] w_i^{(k)} \\ &= \sum_{i=1}^{n} \Psi_i \varepsilon_i(\boldsymbol{\theta}) w_i^{(k)}\end{aligned}$$

with $\varepsilon_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{1}\!\!1(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0) - I\!P(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0)$. Obviously $\mathbb{1}\!\!1(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0)$ is a Bernoulli random variable with the parameter $p_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} I\!P(Y_i - \Psi_i^\top \boldsymbol{\theta} < 0)$. For any $0 < \delta < \mathtt{g}_1 < 1$, there exists a constant $\nu_0 \geq 1$ depending on $\mathtt{g}_1$ only such that

$$\begin{aligned}\log I\!E \exp\{\delta \varepsilon_i(\boldsymbol{\theta})\} &= \log\Big[\big\{ \delta(p_i(\boldsymbol{\theta}) - 1) \big\} p_i(\boldsymbol{\theta}) + \exp\{\delta p_i(\boldsymbol{\theta})\} \{1 - p_i(\boldsymbol{\theta})\} \Big] \\ &\leq p_i(\boldsymbol{\theta}) \{1 - p_i(\boldsymbol{\theta})\} \nu_0^2 \delta^2 / 2.\end{aligned}$$

Therefore, it holds for any $\boldsymbol{\gamma} \in I\!R^{p+1}$ and $\rho > 0$ with $\rho |\boldsymbol{\gamma}^\top \Psi_i| \leq \mathtt{g}_1$ that,

$$\begin{aligned}\log I\!E \exp\{\rho \boldsymbol{\gamma}^\top \nabla\zeta(\boldsymbol{\theta})\} &\leq \log I\!E \exp\left\{ \rho \sum_{i=1}^{n} \boldsymbol{\gamma}^\top \Psi_i \varepsilon_i(\boldsymbol{\theta}) w_i^{(k)} \right\} \\ &\leq \sum_{i=1}^{n} \log I\!E \exp\big\{ \rho \boldsymbol{\gamma}^\top \Psi_i \varepsilon_i(\boldsymbol{\theta}) w_i^{(k)} \big\} \\ &\leq \sum_{i=1}^{n} \rho^2 \big| \boldsymbol{\gamma}^\top \Psi_i w_i^{(k)} \big|^2 p_i(\boldsymbol{\theta}) \{1 - p_i(\boldsymbol{\theta})\} \nu_0^2 / 2 \\ &\leq \nu_0^2 \rho^2 \| V_k(\boldsymbol{\theta}) \boldsymbol{\gamma} \|^2 / 2,\end{aligned}$$

where

$$V_k^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} p_i(\boldsymbol{\theta}) \{1 - p_i(\boldsymbol{\theta})\} \Psi_i \Psi_i^\top {w_i^{(k)}}^2.$$

This particularly yields (ED_0) with $V_{0,k}^2 \stackrel{\text{def}}{=} V_k^2(\boldsymbol{\theta}^*)$ and $\mathtt{g} = \mathtt{g}_1 N_k^{1/2}$. Evidently $V_k^2(\boldsymbol{\theta}) \le V_k^2 \stackrel{\text{def}}{=} (1/4)\sum_{i=1}^n \Psi_i \Psi_i^\top {w_i^{(k)}}^2$ for all $\boldsymbol{\theta}$ and $(E\mathtt{r})$ is fulfilled with $\mathtt{g}(\mathtt{r}) = \mathtt{g}_1 N_k^{1/2}$.

Next we check the local condition (ED_1) on the elliptic vicinity $\Theta_k(\mathtt{r}_0)$. Then

$$\nabla\zeta_k(\boldsymbol{\theta}) - \nabla\zeta_k(\boldsymbol{\theta}^*) = \sum_{i=1}^n \Psi_i\{\varepsilon_i(\boldsymbol{\theta}) - \varepsilon_i(\boldsymbol{\theta}^*)\}w_i^{(k)}$$

Assume $\|V_k\|^2 \le \nu_1 \|V_k\|^2$, for ν_1 as a constant.

$$\begin{aligned}
&\log I\!\!E \exp\bigl[\lambda\boldsymbol{\gamma}^\top\bigl\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\bigr\}\bigr] \\
&\le 4\nu_0^2\lambda^2 \max_{i\le n}\bigl|p_i(\boldsymbol{\theta}) - p_i(\boldsymbol{\theta}^*)\bigr| \|V_k\boldsymbol{\gamma}\|^2/2 \\
&\le \varrho(\mathtt{r})\nu_0^2\lambda^2\|V_k\boldsymbol{\gamma}\|^2/2
\end{aligned}$$

with

$$\varrho(\mathtt{r}) \stackrel{\text{def}}{=} 4\nu_1 \max_{i\le n} \sup_{\boldsymbol{\theta}\in\Theta_k(\mathtt{r})} \{p_i(\boldsymbol{\theta}) - p_i(\boldsymbol{\theta}^*)\},$$

also because

$$|p_i(\boldsymbol{\theta}) - p_i(\boldsymbol{\theta}^*)| \le C|\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)| \le CN_k^{-1/2}\|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le CN_k^{-1/2}\mathtt{r}_0,$$

$\boldsymbol{(ED_1)}$ holds.

The $(\mathcal{L}r)$ and $(\mathcal{L}_0)$ Condition

These identifiability conditions will be checked under the measure $I\!\!P_{\boldsymbol{\theta}^*}$ corresponding to the linear quantile function $f(\cdot) = f_{\boldsymbol{\theta}^*}(\cdot)$. It holds

$$\frac{d I\!\!E_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta})}{d\boldsymbol{\theta}} = -\sum_{i=1}^n \Psi_i\bigl\{\tau - I\!\!P\bigl(Y_i - \Psi_i^\top\boldsymbol{\theta} < 0\bigr)\bigr\}w_i^{(k)}$$

and

$$-\frac{d^2 I\!\!E_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta})}{d^2\boldsymbol{\theta}} = \sum_{i=1}^n \Psi_i\Psi_i^\top \pi_i\{\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\}w_i^{(k)} \stackrel{\text{def}}{=} D_k^2(\boldsymbol{\theta}).$$

Recall that $-\nabla I\!\!E_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}^*) = 0$. Now we take Taylor expansion of $-I\!\!E_{\boldsymbol{\theta}^*} L(W, \boldsymbol{\theta}, \boldsymbol{\theta}^*)$, we conclude that, there is $\boldsymbol{\theta}^\circ \in [\boldsymbol{\theta}, \boldsymbol{\theta}^*]$ such that

$$\begin{aligned} -I\!\!E_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*) &= \sum_{i=1}^n |\Psi_i^\top(\boldsymbol{\theta} - \boldsymbol{\theta}^*)|^2 \pi_i\{\Psi_i^\top(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\} w_i^{(k)}/2 \\ &= (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top D_k^2(\boldsymbol{\theta}^\circ)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)/2. \end{aligned}$$

Moreover, $(\mathcal{L}r)$ need to be proved, for any $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ with $\|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| = \mathtt{r}$, it holds $\pi_i\{\Psi_i^\top(\boldsymbol{\theta}^\circ - \boldsymbol{\theta}^*)\} \geq (1-\delta)\pi_i(0)$ (A.4).

$$-I\!\!E_{\boldsymbol{\theta}^*} L(W, \boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \frac{1-\delta}{2} \|D_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \geq \frac{1-\delta}{2\mathtt{a}^2} \|V_k(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 = \frac{1-\delta}{2\mathtt{a}^2} \mathtt{r}^2.$$

This easily yields

$$-I\!\!E_{\boldsymbol{\theta}^*} L(W^{(k)}, \boldsymbol{\theta}, \boldsymbol{\theta}^*)/\mathtt{r}^2 \geq \frac{1-\delta}{2\mathtt{a}^2}$$

So, the global identifiability condition $\boldsymbol{(\mathcal{L}r)}$ is fulfilled if $\mathtt{r}^2 \geq C_1(\mathtt{x} + p)$ for some fixed constants C_1 . Also, as $D^2(\boldsymbol{\theta}^\circ) \geq (1-\delta) D_k^2$ leads to $\|I_{p+1} - D_k^{-1} D^2(\boldsymbol{\theta}^\circ) D_k^{-1}\|_\infty \leq \delta$ for $\boldsymbol{\theta} \in \Theta_k(\mathtt{r})$, $(\mathcal{L}_0)$ holds.

2.7.3 Theorem for critical values

The theorem below assures an upper bound for the critical values $\mathfrak{z}_k$ constructed in Section 2.2.5. To avoid technical burdens, we restrict the analysis to the random set $\mathcal{A}$ and discard the large deviation probability part on its complement. The notation $I\!\!P'(B)$ for a set B means $I\!\!P(B \cap \mathcal{A})$.

Theorem 2.7.4. *Suppose that $r > 0, \alpha > 0$. There exist constants a_0, a_1 s.t. the propagation condition is fulfilled with the choice of*

$$\mathfrak{z}_k = a_0 + \log(\alpha^{-1}) + a_1 r(K-k) + r\log(p) \tag{2.24}$$

Proof. First we bound the quantity $L\big(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\big)$ on the random set $\mathcal{A} = \mathcal{A}_1 \cap \mathcal{A}_2$. The majorization (2.21) and its corollary (2.22) yield on $\mathcal{A}$ with $\boldsymbol{u}_{\ell k} \stackrel{\text{def}}{=}$

$D_{\underline{\boldsymbol{\epsilon}},k}\bigl(\widetilde{\boldsymbol{\theta}}_\ell(x) - \boldsymbol{\theta}^*\bigr)$

$$\begin{aligned} L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) &= L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \boldsymbol{\theta}^*\bigr) - L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_\ell(x), \boldsymbol{\theta}^*\bigr). \\ &\le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2 - \mathbb{L}_{\underline{\boldsymbol{\epsilon}}}\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_\ell(x), \boldsymbol{\theta}^*\bigr) + \diamondsuit_{\boldsymbol{\epsilon},k} \\ &= \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2/2 - \boldsymbol{u}_{\ell k}^\top \boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}},k} + \|\boldsymbol{u}_{\ell k}\|^2/2 + 2\diamondsuit_{\boldsymbol{\epsilon},k} \\ &\le \bigl(\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\| + \|\boldsymbol{u}_{\ell k}\|\bigr)^2/2 + 2\diamondsuit_{\boldsymbol{\epsilon},k} \\ &\le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2 + \|\boldsymbol{u}_{\ell k}\|^2 + 2\diamondsuit_{\boldsymbol{\epsilon},k}, \end{aligned} \tag{2.25}$$

where we used the fact that $\|\boldsymbol{\xi}_{\underline{\boldsymbol{\epsilon}},k}\| \le \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|$.

It is not difficult to see that

$$\|\boldsymbol{u}_{\ell k}\|^2 = \|D_{\underline{\boldsymbol{\epsilon}},k} D_{\boldsymbol{\epsilon},\ell}^{-1} D_{\boldsymbol{\epsilon},\ell}\bigl(\widetilde{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}^*\bigr)\|^2 \le \|D_{\underline{\boldsymbol{\epsilon}},k} D_{\boldsymbol{\epsilon},\ell}^{-2} D_{\underline{\boldsymbol{\epsilon}},k}\|_\infty \, \|D_{\boldsymbol{\epsilon},\ell}\bigl(\widetilde{\boldsymbol{\theta}}_\ell - \boldsymbol{\theta}^*\bigr)\|^2.$$

By construction $D^2_{\boldsymbol{\epsilon},k} \le D^2_k \le D^2_{\underline{\boldsymbol{\epsilon}},k}$ and the definition (2.20) implies by $\alpha_{\boldsymbol{\epsilon},k} \le 1/2$

$$D^2_{\underline{\boldsymbol{\epsilon}},k} \le (1 - \alpha_{\boldsymbol{\epsilon},k})^{-1} D^2_{\boldsymbol{\epsilon},k} \le 2 D^2_{\boldsymbol{\epsilon},k}.$$

Now it follows from condition A.2 that

$$\|D_{\underline{\boldsymbol{\epsilon}},k} D_{\boldsymbol{\epsilon},\ell}^{-2} D_{\underline{\boldsymbol{\epsilon}},k}\|_\infty \le 2\|D_k D_\ell^{-2} D_k\|_\infty \le \begin{cases} 2/\mathfrak{u}_0^{k-\ell}, & k > \ell, \\ 2\mathfrak{u}^{\ell-k}, & k < \ell. \end{cases} \tag{2.26}$$

Corollary 2 implies

$$\|D_{\boldsymbol{\epsilon},\ell}\bigl(\widetilde{\boldsymbol{\theta}}_\ell(x) - \boldsymbol{\theta}^*\bigr)\| \le 2\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + \bigl(1 + \alpha_{\boldsymbol{\epsilon},\ell}^{1/2}\bigr)\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\| \le 2\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + 2\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\|. \tag{2.27}$$

We also use that $I\!\!E_{\boldsymbol{\theta}^*}\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^{2r} \le p^r C_r(\mathrm{A})$ for all $k \le K$. Now it holds from (2.25), (2.26), and (2.27) for $k > \ell$

$$\begin{aligned} I\!\!E'_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) &\le I\!\!E'_{\boldsymbol{\theta}^*}\Bigl[\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2 + 8\mathfrak{u}_0^{-k+\ell}\bigl(\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\|\bigr)^2 + 2\diamondsuit_{\boldsymbol{\epsilon},k}\Bigr]^r \\ &\le C(\mathrm{A}) p^r \mathfrak{u}_0^{-r(k-\ell)}. \end{aligned} \tag{2.28}$$

Similarly one can show that for $k < \ell$ by $\mathfrak{u} < 1$

$$\begin{aligned} I\!\!E'_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) &\le I\!\!E'_{\boldsymbol{\theta}^*}\bigl[\|\boldsymbol{\xi}_{\boldsymbol{\epsilon},k}\|^2 + 8\bigl(\diamondsuit_{\boldsymbol{\epsilon},\ell}^{1/2} + \|\boldsymbol{\xi}_{\boldsymbol{\epsilon},\ell}\|\bigr)^2 + 2\diamondsuit_{\boldsymbol{\epsilon},k}\bigr]^r \\ &\le C(\mathrm{A}) p^r. \end{aligned}$$

Also by Theorem 2.7.3 for $\mathtt{x} > 0$

$$\mathbb{P}_{\boldsymbol{\theta}^*}\bigl\{L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr) > C_1 p + C_2 \mathtt{x}\bigr\} \le 2\mathrm{e}^{-\mathtt{x}}. \tag{2.29}$$

These bounds can be used to check that the critical value $\mathfrak{z}_k$ which is selected in the form (2.24) to ensure the propagation condition in (2.10). Consider a random set $\mathcal{B}_\ell \stackrel{\text{def}}{=} \{\widehat{k}(x) = \ell\}$, By definition of $\widehat{k}$, when $\mathcal{B}_\ell$ happens, at least one of the estimate $\widetilde{\boldsymbol{\theta}}_{\ell+1}(x)$ must be not accepted, that is,

$$\mathcal{B}_\ell \subseteq \bigcup_{m=1}^{\ell} \Bigl\{L\bigl(W^{(m)}, \widetilde{\boldsymbol{\theta}}_m(x), \widetilde{\boldsymbol{\theta}}_{\ell+1}(x)\bigr) > \mathfrak{z}_m\Bigr\}.$$

The bounds (2.28) and (2.29) yield by the Cauchy-Schwarz inequality

$$\begin{aligned}
&\mathbb{E}'_{\boldsymbol{\theta}^*} L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr) \\
&\quad\le \sum_{\ell=1}^{k} \bigl[\mathbb{E}'_{\boldsymbol{\theta}^*} L^{2r}\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widetilde{\boldsymbol{\theta}}_\ell(x)\bigr)\bigr]^{1/2} \bigl[\mathbb{P}'_{\boldsymbol{\theta}^*}(\mathcal{B}_\ell)\bigr]^{1/2} \\
&\quad\le C(\mathrm{A}) p^{2r} \sum_{\ell=1}^{k} \mathfrak{u}_0^{-2r(k-\ell)} \bigl[\mathbb{P}'_{\boldsymbol{\theta}^*}(\mathcal{B}_\ell)\bigr]^{1/2} \\
&\quad\le C(\mathrm{A}) p^{2r} \sum_{\ell=2}^{k} \mathfrak{u}_0^{-2r(k-\ell)} \biggl[\sum_{m=1}^{\ell} \mathbb{P}'_{\boldsymbol{\theta}^*}\Bigl\{L\bigl(W^{(m)}, \widetilde{\boldsymbol{\theta}}_m(x), \widetilde{\boldsymbol{\theta}}_{\ell+1}(x)\bigr) > \mathfrak{z}_m\Bigr\}\biggr]^{1/2}.
\end{aligned}$$

Fix $c_0 > \log(\mathfrak{u}_0^{-1})$ and consider $\mathfrak{z}_m = C_1 p + C_2 \mathtt{x}_m$ with $\mathtt{x}_m = 2c_0 r(K-m) + 2\mathtt{x}$ for some $\mathtt{x}$. Then (2.29) implies

$$\begin{aligned}
&\mathbb{E}'_{\boldsymbol{\theta}^*}\bigl[L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr)\bigr] \\
&\quad\le C(\mathrm{A}) p^{2r} \sum_{\ell=2}^{K} \mathfrak{u}_0^{-2r(K-\ell)} \biggl[\sum_{m=1}^{\ell} 2\mathrm{e}^{-\mathtt{x}_m}\biggr]^{1/2} \\
&\quad\le C(\mathrm{A}) p^{2r} \mathrm{e}^{-\mathtt{x}} \sum_{\ell=2}^{K} \exp\bigl[-2r(K-\ell)\bigl\{c_0 - \log(1/\mathfrak{u}_0)\bigr\}\bigr] \\
&\quad\le C(\mathrm{A}) p^{2r} \mathrm{e}^{-\mathtt{x}}
\end{aligned}$$

and the bound (2.10) follows with $\mathtt{x} = \log(1/\alpha) + r\log(p) + a_0$ for a proper a_0. $\square$

2.7.4 Propagation Property and Stability

The "oracle" result is a consequence of two properties of the procedure: "propagation" under homogeneity and "stability"; see Appendix for a precise definition. The first one means that the procedure would not terminate for $k < k^*$ (no false alarm) with a high probability. The "stability" property ensures that the estimation quality will not essentially deteriorate in the steps after "propagation" for $k > k^*$. By construction, the procedure described in Section 2 provides the prescribed performance if the true signal $f(\cdot)$ follows the parametric model (local constant or local linear). Now, the following theorem implies similar performance under the true nonparametric model $f(\cdot)$ before the oracle k^*. From the above lemma, we can derive the propagation property from the propagation condition:

Theorem 2.7.5. *Assume the SMB condition $\Delta_{k^*}(\boldsymbol{\theta}) \le \Delta$ for some k^*. Then*

$$\mathbb{E} \log\bigl\{1 + L^r\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widehat{\boldsymbol{\theta}}_{k^*}(x)\bigr)/\mathcal{R}_r\bigr\} \le \Delta + \alpha, , \tag{2.30}$$

$$\mathbb{E} \log\bigl\{1 + L^r\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \boldsymbol{\theta}\bigr)/\mathcal{R}_r\bigr\} \le \Delta + 1. \tag{2.31}$$

The result (2.31) shows that the estimation loss $L^r\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k(x), \widehat{\boldsymbol{\theta}}_k(x)\bigr)$ normalized by the parametric risk bound $\mathcal{R}_r$ is stochastically bounded by a constant of order e^{Δ}.

Due to the "propagation" result (2.10), the accuracy of the sequential test is guaranteed when the SMB assumption is fulfilled. In addition, we also need to make sure that when our final estimated step $\widehat{k}(x)$ overshoots the oracle $k^*(x)$, that is, $\widehat{k}(x) > k^*(x)$, the estimators $\widetilde{\boldsymbol{\theta}}_k$ does not vary too much. The stability property can be stated as follows.

Theorem 2.7.6. *In the case of overshooting $\widehat{k} > k^*(k^* = k^*(x))$, the estimate $\widehat{\boldsymbol{\theta}}(x)$ fulfills*

$$L\bigl(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widehat{\boldsymbol{\theta}}(x)\bigr)\, 1\!\!1\{\widehat{k}(x) > k^*\} \le \mathfrak{z}_{k^*}.$$

This assertion follows from the setup of our test because the estimate $\widehat{\boldsymbol{\theta}}(x) = \widetilde{\boldsymbol{\theta}}_{\widehat{k}(x)}(x)$ is accepted and for $\widehat{k} > k^*$, it should be in the confidence set of $\widetilde{\boldsymbol{\theta}}_{k^*(x)}(x)$.

2.7.5 Proof of the "oracle" property

Lemma 2. *Let P, P_0, be two measures s.t. $\mathbb{E} \log(dP/dP_0) \le \Delta < \infty$. For any random variable Z with $\mathbb{E}Z < \infty$, it holds $\mathbb{E} \log(1 + Z) \le \Delta + \mathbb{E}_0 Z$.*

Proof. The function $f(x) = xy - x\log x + x$ attains maximum at the point $x = e^y$, thus $f(x) \le f(e^y)$, and thus $xy \le x\log x - x + e^y$. With $X = d\mathbb{P}/d\mathbb{P}_0$ and $Y = \log(1+Z)$, it holds

$$\begin{aligned}
\mathbb{E}\log(1+Z) &= \mathbb{E}_0\big\{X\log(1+Z)\big\} \\
&\le \mathbb{E}_0(X\log X - X + 1 + Z) \\
&\le \mathbb{E}\log\frac{d\mathbb{P}}{d\mathbb{P}_0} + \mathbb{E}_0 Z \le \Delta + \mathbb{E}_0 Z
\end{aligned}$$

as required. □

Proof. (2.15) is a trivial consequence of (2.30) and "stability". We now prove (2.16).

$$\begin{aligned}
&\mathbb{E}\log\left\{1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right\} \\
&\quad = \mathbb{E}\left[\log\left\{1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right\} \mathbb{1}\!\mathrm{I}(\widehat{k} \le k^*)\right] \\
&\qquad + \mathbb{E}\left[\log\left\{1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right\} \mathbb{1}\!\mathrm{I}(\widehat{k} > k^*)\right] \\
&\quad \le \Delta + \mathbb{E}\left[\frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\right] + \mathbb{E}\log\left[1 + \frac{L^r\big(W^{(k^*)}, \widetilde{\boldsymbol{\theta}}_{k^*}(x), \widetilde{\boldsymbol{\theta}}_{\widehat{k}}(x)\big)}{\mathcal{R}_r}\,\mathbb{1}\!\mathrm{I}(\widehat{k} > k^*)\right] \\
&\quad \le \Delta + \rho + \log(1 + \mathfrak{z}_{k^*}/\mathcal{R}_r)
\end{aligned}$$

Chapter 3

Tie the straps: uniform bootstrap confidence interval for additive models

3.1 Introduction

We consider in this chapter conditional M- and L- estimates with regressors $X \in \mathbb{R}^d$. The set of estimators includes in particular conditional quantiles and bounded influence smoothers. For d-dimensional regressors X, we run of course into a dimensionality problem. One way to avoid this problem is to impose a simplified structure (such as additive) on the multivariate nonparametric function. The additive structure assumes that the covariates' effects are separable, and this effect is presented in many economic applications, Härdle (1990). Specifically, the structure considered is:

$$m(x_1, \ldots, x_d) = \sum_{j=0}^{d} m_j(x_j), \tag{3.1}$$

with $m_0(x_j)$ a constant. It is well known that (3.1) achieves dimension reduction in the sense that one dimensional convergence rates are achieved for approximation of $m(x_1, \ldots, x_d)$ in (3.1), see Horowitz & Lee (2005) for quantile regression and additive modeling.

The additional contribution from our results is that the bootstrap based confidence bands are shown to be very close to the finite sample distribution based ones. Our construction is also applicable to confidence bands where we obtain surprisingly precise approximation to the randomness of the bands.

Additive modeling is an important way to achieve dimension reduction in multivariate regression, e.g. Horowitz (2001b), Horowitz & Lee (2005), Horowitz, Klemelä & Mammen (2006), and among many others. Fully nonparametric smoothing is non attractive in high dimension because the curse of dimensionality causes imprecision for data sizes typically found in applications. Nonparametric additive models reduce this imprecision problem to rates of convergence typical for one dimensional regression and still provide flexibility of marginal influence (i.e. the effect of X_i on Y_i). The resulting estimate $\widehat{m}_j(x_j)$ in (3.1) though needs to be screened for closeness to $m_j(x_j)$. This requires construction of confidence intervals and bands (as a function of x_j). For such screening tests, our tightened bootstrap techniques will be verified.

The bootstrap is a class of data driven sampling techniques that provide nonasymptotic approximations of the finite sample distribution of different statistics. In a location model (more generally a regression model), resampling is done from the estimated residuals and typical theoretical analysis leads to the conclusion "bootstrap works" in the sense that the suitably centered bootstrap estimator converges to the same asymptotic normal distribution as the original estimator under consideration. A large literature body has focused on showing bootstrap improvements and refinements of approximations via bootstrap resampling, see Hall (1992), Mammen (1992), Horowitz (2001a), Hardle, Horowitz & Kreiss (2003), which discuss the conditions for bootstrap consistency, and also prove the bootstrap accuracy as an approximation to the exact finite sample distribution for special types of statistics in a nonparametric framework. But very few of them has been focused on nonlinear statistics (e.g. maximum) in nonparametric regression, because it is difficult to analyze the bootstrap improvement in this case.

In this chapter, we investigate a coupling technique that allows us to "tie the straps" even a little tighter for a class of estimators. We mean by that, theoretically speaking, confidence interval construction is made more precise in a variety of the estimation problems we consider for the generalized linear models. The coupling idea is based on mimicking the distribution of the original data via a controllable random mechanism.

Let us describe the coupled bootstrap in the simple case of nonparametric quantile framework as in Härdle, Ritov & Song (2012). Here $(X, Y)^\top \in \mathbb{R}^2$ and $l(x) = F^{-1}_{(Y|X=x)}(\tau)$ is the conditional $\tau-$quantile function. The choice for $\tau = 1/2$ yields the conditional median regression. The conditional quantile function can be attained by:

$$l(x) = \arg\min_\theta I\!E_{(Y|X=x)}\rho(Y-\theta), \tag{3.2}$$

where $\rho(u) = \tau u\mathbf{f}(u > 0) - (1-\tau)u\mathbf{f}(u < 0)$ is the check function for τ- th quantile. Alternatively, (3.2) may be seen as a minimum contrast parameter based on the

log likelihood function of an $ALD(\tau)$ (asymmetric Laplace distribution). A sample based estimator of (3.2) is:

$$\widehat{l}_h(x) = \text{argmin}_\theta n^{-1} \sum_{i=1}^{n} \rho(Y_i - \theta) K_h(x - X_i), \tag{3.3}$$

where $K_h(u) = K(u/h)/h$ is a kernel function with bandwidth h. Note that (3.3) is typically the estimation under conditional location model.

One can generate a bootstrap sample using an i.i.d. standard uniform random variables $U_1, \ldots, U_n$, and then generate:

$$Y_i^* = \widehat{l}_g(X_i) + \varepsilon_i^*, \quad i = 1, \ldots, n, \tag{3.4}$$

where $\varepsilon_i^* = \widehat{F}^{-1}_{(Y|X=x_i)}(U_i)$ and g a slightly larger bandwidth than h. The basic idea of coupling is based on comparing this sample to the pseudo observations:

$$Y_i^\sharp = l(X_i) + \varepsilon_i^\sharp, \quad i = 1, \ldots, n, \tag{3.5}$$

where $\varepsilon_i^\sharp = F^{-1}_{(Y|X=x_i)}(U_i)$. Note that given $\{X_i\}_{i=1}^n$, the distribution of $Y_i^\sharp$ and Y_i are the same. We will show for a class of loss functions that the following strong approximation holds:

$$\sup_{x \in B} \left[\widehat{l}_h^\sharp(x) - l(x) - \{\widehat{l}^*_{h,g}(x) - \widehat{l}_g(x)\} \right] = \mathcal{O}(h^2 \Gamma_n), \tag{3.6}$$

where Γ_n a slowly varying sequence (a sequence a_n is slowly varying if $n^{-\alpha} a_n \to 0$ for any $\alpha > 0$), $\widehat{l}_h^\sharp(\cdot)$ is the nonparametric estimate calculated from $\{(X_i, Y_i^*)\}$, $\widehat{l}^*_{h,g}(X_i)$ is an estimate calculated from the bootstrap sample $\{(X_i, Y_i^*)\}$ with bandwidth h, where

$$\widehat{\ell}^*_{h,g} \stackrel{\text{def}}{=} \text{argmin}_\theta \sum_{i=1}^{n} \rho(Y_i^* - \theta) K_h(x - X_i) \tag{3.7}$$

and $\widehat{l}_g(X_i)$ is calculated as in (3.3) from the original sample with bandwidth g.

The basic elements in proving (3.6) are smoothness of $F_{Y|X=x}(\cdot)$ and bounded influence of $\rho(\cdot)$ in (3.2). Similar results like (3.6) will be derived for additive models. Additive modeling is an important way to achieve dimension reduction in multivariate regression: Horowitz (2001b) focuses on generalized additive models with unknown link functions, Horowitz & Lee (2005) propose a two-stage estimation for quantile regression in additive models, Horowitz et al. (2006) show the equivalence between spline, kernel and other methods in terms of optimal minimax rate in additive model estimation. Fully nonparametric smoothing is non attractive in high dimension because the curse of dimensionality causes imprecision for data

sizes typically found in applications. Nonparametric additive models reduce this imprecision problem to rates of convergence typical for one dimensional regression and still provide flexibility of marginal influence (i.e. the effect of $X_{i,j}$ on Y_i). The resulting estimate $\widehat{m}_j(x_j)$ in (3.1) though needs to be screened for closeness to $m_j(x_j)$. This requires construction of confidence intervals and bands (as a function of x_j). For such screening tests, our tightened bootstrap techniques will be verified. The remainder of the chapter is organized as follows. In Section 3.2 we explain in details the model setup and the bootstrap method. Section 3.3 presents the main results. In Section 3.4 a small simulation study is presented. Finally, we show in Section 3.5 some applications.

3.2 Additive models and bootstrap confidence sets

This section describes our coupling techniques informally, motivates the obtainable theoretical results and discusses some of the assumptions. For any $x \in \mathbb{R}^d$, define first the nonparametric M-estimate

$$\widehat{l}_h(x) = \arg\min_\theta \sum_{i=1}^{n} \rho(Y_i - \theta) K_h(x - X_i), \tag{3.8}$$

with a (d-dimensional) kernel $K_h(\cdot)$. The estimator $\widehat{l}_h(x)$ will estimate the minimum contrast parameter:

$$l(x) = \mathrm{argmin}_\theta I\!E_{(Y|X=x)}\{\rho(Y - \theta)\} \tag{3.9}$$

Here $\rho(\cdot)$ is a loss function of Hampel/Huber type or more generally (up to a constant) a negative (pseudo) log likelihood. In the quantile regression case, $\rho(x) = x\{\tau - \mathbf{f}(x \le 0)\}$ is the check function . An example for $\rho(\cdot)$ is a trimmed mean, Huber (1964)

$$\rho(x) = \begin{cases} x^2, & |x| \le k, \\ k^2, & |x| > k \end{cases}, \tag{3.10}$$

or a form of Winsorized mean:

$$\rho(x) = \begin{cases} x^2/2, & |x| \le k, \\ -k^2/2 + k|x|, & |x| > k. \end{cases} \tag{3.11}$$

The nonparametric approach in (3.8) is not appropriate when d is large. Typically, the optimal convergence rate $\mathcal{O}(n^{-4/(4+d)})$ would be slower when d is large. Additive models were suggested to remedy the problems posed by the dimension. Recall

in (3.1), without abusing of notations, write the multivariate additive function as $m(\cdot)$:

$$Y_i = m(X_i) + \varepsilon_i \tag{3.12}$$

where

$$m(X_i) = \sum_{j=0}^{d} m_j(x_{i,j}). \tag{3.13}$$

We further approximate the additive model via a basis function approach:

$$m_j(x_{i,j}) \approx \sum_{l=1}^{L_j+1} a_{l,j} g_l(x_{i,j}),$$

where the $g_1(\cdot), g_2(\cdot), \ldots$ could be any sequence of functions spanning L_2. Our implementation uses the B-splines, for example, linear B-splines: Consider a sequence of H equally spaced knots on the interval $[0,1]$, which defines the width H^{-1} subintervals $[lH^{-1}, (l+1)H^{-1}]$, $0 \le l \le N+1$. For $l = 0, \ldots, H$. The linear B-spline basis is:

$$g_l(x) = \begin{cases} Hx - l + 1 & (l-1)H^{-1} \le x \le lH^{-1} \\ l + 1 - Hx & lH^{-1} \le x \le (l+1)H^{-1} \\ 0 & \text{otherwise} \end{cases}$$

Denote the theoretical standardized B spline basis $\phi_l(\cdot)$,

$$\begin{aligned} \phi_{l,j}(x_j) &= g_l(x_j) - g_{l-1}(x_j) c_{l,j}/c_{l-1,j} \\ B_{j,l}(x_j) &= \phi_{l,j}(x_j)/\|\phi_{l,j}(x_j)\|_2, 1 \le J \le H, \end{aligned}$$

where $l = 0, \ldots, H$, $c_{l,j} = \int \phi_{l,j}(x_j) f_j(x_j) dx_j$, so that $I\!E B_{j,l}(x_j) = 0$, $I\!E B_{j,l}(x_j)^2 = 1$.
Therefore, similar to (3.8), the additive estimate can be obtained. Define the following vectors in $\mathbb{R}^{Ld+1}$

$$\begin{aligned} A &= (a_0, \mathbf{a}_1^\top, \ldots, \mathbf{a}_d^\top)^\top \\ \Phi(X_i) &= \{1, \mathbf{g}(x_{i,1})^\top, \ldots, \mathbf{g}(x_{i,d})^\top\}^\top, \end{aligned}$$

where

$$\begin{aligned} \mathbf{a}_j &= (a_{1,j}, \ldots, a_{(L_j+1),j})^\top \\ \mathbf{g}(x_{i,j})^\top &= \{g_1(x_{i,j}), \ldots, g_H(x_{i,j})\}^\top. \end{aligned}$$

Finally, let $\widehat{A}_{\mathbb{L}}$ be the estimation of A:

$$\widehat{A}_{\mathbb{L}} = \arg\min_A \sum_{i=1}^{n} \rho\{Y_i - A^\top \Phi(X_i)\}. \tag{3.14}$$

3.2.1 Coupled Bootstrap for Quantiles

The additive structure in (3.13) is one solution to the curse of dimensionality problem, however, the bootstrap approach in (3.6) does not work for this modeling scenario. We suggest another bootstrap technique, and prove that it strongly approximate a model with the same asymptotic properties as the original model.

Define

$$Z_i = \begin{cases} 1 & \text{with prob } \tau \\ -1 & \text{with prob } 1-\tau \end{cases}, \quad i = 1, \ldots, n. \tag{3.15}$$

The bootstrap couple ε^* (the bootstrap residuals) and $\varepsilon^\sharp$ (the theoretical couple) are:

$$\varepsilon^* \stackrel{\text{def}}{=} Z_i |\widehat{\varepsilon}_i| \tag{3.16}$$

$$\varepsilon^\sharp \stackrel{\text{def}}{=} Z_i \eta_i, \quad i = 1, \ldots, n, \tag{3.17}$$

where

$$F_{i,s}(t) \stackrel{\text{def}}{=} I\!P(|\varepsilon_i| \leq t | s\varepsilon_i > 0), \quad i = 1, \ldots, n, \ s \in \{1, -1\}, \tag{3.18}$$

and

$$\eta_i \stackrel{\text{def}}{=} F^{-1}_{i,Z_i}\{F_{i,\text{sgn}(\varepsilon_i)}(|\varepsilon_i|)\}, \quad i = 1, \ldots, n. \tag{3.19}$$

Recall that $F_{Y|X=x_i}\{l(X_i)\} = \tau$ and $F_{\varepsilon|X=x_i}(0) = \tau$. Now, it is easy to see that $V_i \stackrel{\text{def}}{=} F_{i,\text{sgn}(\varepsilon_i)}(|\varepsilon_i|)$ has a standard uniform distribution, and if Z_i is as above, then ε_i and $Z_i F^{-1}_{i,Z_i}(V_i)$ have the same distribution. Formally, note that

$$F_{i,+1}(t) = \frac{F_i(t) - 1 + \tau}{\tau},$$
$$F_{i,-1}(t) = \frac{1 - \tau - F_i(-t)}{1 - \tau},$$

where $F_i(\cdot)$ is the cdf of ε_i.

Hence, for $t > 0$:

$$\begin{aligned}
I\!P(\varepsilon_i^\sharp < t) &= \tau I\!P[F_{i,+1}^{-1}\{F_{i,\mathrm{sgn}(\varepsilon_i)}(|\varepsilon_i|)\} < t] + 1 - \tau \\
&= \tau I\!P\{F_{i,\mathrm{sgn}(\varepsilon_i)}(|\varepsilon_i|) < F_{i,+1}(t)\} + 1 - \tau \\
&= \tau I\!P\{\varepsilon_i < 0, F_{i,-1}(-\varepsilon_i) < F_{i,+1}(t)\} \\
&\qquad + \tau I\!P\{\varepsilon_i > 0, F_{i,+1}(\varepsilon_i) < F_{i,+1}(t)\} + 1 - \tau \\
&= \tau I\!P\{\varepsilon_i < 0, \frac{1-\tau-F_i(\varepsilon_i)}{1-\tau} < \frac{F_i(t)-1+\tau}{\tau}\} \\
&\qquad + \tau I\!P(0 < \varepsilon_i < t) + 1 - \tau \\
&= \tau I\!P[1-\tau > F_i(\varepsilon_i) > \frac{1-\tau}{\tau}\{1-F_i(t)\}] \\
&\qquad + \tau I\!P(0 < \varepsilon_i < t) + 1 - \tau \\
&= \tau[1 - \frac{1-\tau}{\tau}\{1-F_i(t)\} - \tau] + \tau\{F_i(t) - 1 + \tau\} + 1 - \tau \\
&= F_i(t).
\end{aligned}$$

The case $t < 0$ is dealt similarly. It follows

$$\mathcal{L}(\varepsilon_i^\sharp) = \mathcal{L}(\varepsilon_i). \tag{3.20}$$

Our confidence "ideal" interval is conditional on $\{V_i\}_{i=1}^n$ which has a direct link to the absolute value of the residuals $\{|\varepsilon_i|\}_{i=1}^n$. Note however that the estimator is asymptotically consistent and its bias does not depend on these absolute values. Moreover, by the law of large numbers, the pointwise width of the conditional confidence interval is within a factor of $1 + \mathcal{O}_p(1)$ of the unconditional one.

3.2.2 How does the coupling work?

The basic idea of our approach, is trying to construct an empirically feasible bootstrap sample that is strongly approximating a sample from the true distribution. One example of the coupled bootstrap sample was already explained in (3.5) and (3.4). It however relies on estimators of the conditional distribution $F_{Y|X=x}(.)$, which become very imprecise when d is large.

Another approach motivated as the wild bootstrap is based on randomizing the obtained residuals and using the same random source to mimic the stochastic of the unobservable errors. To get the basic idea, let us assume for a moment that the distributions of ε_i are symmetric. Then the coupling may be performed via a Rademacher randomized variables Z_i with

$$I\!P(Z_i = 1) = I\!P(Z_i = -1) = 1/2$$

and generation of the couple ε_i^* (the bootstrapped residuals), $\varepsilon_i^\sharp$ (the theoretical coupling), where $\{\varepsilon_i^*, \varepsilon_i^\sharp\}$ is :

$$\begin{aligned} \varepsilon_i^* &\stackrel{\text{def}}{=} Z_i \widehat{\varepsilon}_i \\ \varepsilon_i^\sharp &\stackrel{\text{def}}{=} Z_i \eta_i. \end{aligned} \tag{3.21}$$

With this construction, we are able to establish a result similar to (3.6).

In a non symmetric distribution (required for quantile regression), one defines Z_i with $I\!P(Z_i = 1) = \tau$ and $I\!P(Z_i = -1) = 1 - \tau$ assuming the centering $F_{Y|X_i}\{l(X_i)\} = \tau$, and the couple $(\varepsilon_i^*, \varepsilon_i^\sharp)$ is given by (3.17) and (3.16). It was argued that the distributions of $\varepsilon_i^\sharp$ and ε_i are identical and also the conditional distributions given $\{V_i\}_{i=1}^n$ are the same.

The resampling technique will be applied to nonparametric estimation of an additive quantile regression model. The reanalysis of the Data used by Horowitz & Lee (2005) provides us with sharper bands that have not been calculated in that chapter.

3.3 Main Results

The section gives asymptotic results for the estimators described in Section 3.2. To establish the asymptotic property, some assumptions are needed:

Assumptions

A.1 The function $l(x)$ solves (3.9) and it is four-times differentiable, also (3.14) with $\psi(.) = \rho'(.)$ being a.s. differentiable and Lipschitz continuous: $\forall \mu_1, \mu_2 \in B$ (suppose B is the compact set that $l(x)$ takes value on), $|\psi(\mu_1) - \psi(\mu_2)| < C|\mu_1 - \mu_2|$, and we assume that $\exists M > 0$ s.t.$\psi(\mu) \leq M$.

A.2 Assume the support of X is $[0,1]^d$. The conditional density $f_{(\varepsilon|X=x)}(\cdot)$ is bounded from below in the sense that $\forall$ small constant $b > 0$ exists C_1, c_1 such that $\infty > C_1 > \inf_{t\in[-b,b]} f_{(\varepsilon|X=x)}(t) = c_1 > 0$.

A.3 The kernel function $K(\cdot)$ is a product kernel composed from one dimension kernel with bandwidth $h = h_n$:

$$K_h(s) = \Pi_{j=1}^d K(s_j/h)/h, s = (s_1, \ldots, s_d)^\top \in \mathbb{R}^d. \tag{3.22}$$

A.4 The bandwidth satisfies $h \sim n^{-1/(4+d)}$. Let g be another bandwidth sequence $g >> h$. $\{g = \mathcal{O}(n^{-1/9})\}$. Let Γ_n be a slowly increasing sequence in the sense that $n^{-\alpha}\Gamma_n \to 0$ for any $\alpha > 0$.

A.5 For each j, $m_j(\cdot)$, $j \in 1, \ldots, d$, is a at least one time continuous differentiable function on $[0,1]$, and there is an $\alpha > 0$ such that

$$\begin{aligned} I\!E\{\sum_{i=1}^{d} m_j(X_j)\}^2 &\geq \alpha \max_j I\!E\{m_j^2(X_j)\} \\ I\!E\{m_j(X_j)\} &= 0, \end{aligned}$$

for any $m_j(\cdot) \in L_2(X_j)$.

A.6 $I\!E\{g_l^2(x_{i,j})\} = 1$ for any $i \in 1, \ldots, n$ and $j \in 1, \ldots, d$. $||\Phi_l(X_j)||_\infty \leq C_3/L, a.s.$, where $\Phi_l(X_j) \stackrel{\text{def}}{=} \{\phi_l^2(x_{1,j}), \ldots, \phi_l^2(x_{n,j})\}^\top$, with $j \in 1, \ldots, d$.

A.7 The number of regressors in (3.1) is of order $p = dL + 1$ with $L \sim n^{1/5}$.

A.1 is about the continuity and the bounded influence structure of the loss function, we believe that it is quite essential for proving the bootstrap improvement. A.2 and A.3 are assumptions on conditional density and the kernel function. A.4 is about the oversmoothing idea to improve the bootstrap performance, see Härdle & Marron (1991).

We prove first convergence results for bootstrap method in (3.4) and (3.5) . The resampling step has been defined in (3.4), where the smooth estimate of the conditional distribution is:

$$\widehat{F}_{(Y|X=x)}(t) = \sum_{i=1}^{n} W_{h,i}(x) \mathbf{f}(Y_i \leq t), \tag{3.23}$$

with $W_{h,i}(x) = n^{-1} K_h(x - X_i)/\widehat{f}_h(x)$ and $\widehat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$ the well known kernel density estimator.

Once Y_i^* are generated, one applies (3.8) to the bootstrap data $\{(X_i, Y_i^*)\}_{i=1}^n$ to obtain $\widehat{l}_{h,g}^*(x)$.

Theorem 3.3.1. *Let assumptions A.1- A.4 be fulfilled and define $\widehat{l}_h(\cdot)$ as in (3.8). Then*

$$\sup_{x \in B} |A_n(x)| = \mathcal{O}_{p(h^2 \Gamma_n),} \tag{3.24}$$

where

$$A_n(x) \stackrel{\text{def}}{=} (\widehat{l}_h - l)(x) - \{(\widehat{l}_{h,g}^* - \widehat{l}_g)(x)\}. \tag{3.25}$$

Remark 1 If the influence function proportional to $\psi(\cdot) = \rho'(\cdot)$ of the estimator is bounded with bounded derivatives a.e. and a consistent estimator of the conditional distribution $\mathcal{L}(\varepsilon|X)$ exists with $||\widehat{F}_{(\varepsilon|x=X_i)}(\cdot) - F_{(\varepsilon|X=x_i)}(\cdot)||_\infty = \mathcal{O}_{(h^2\Gamma_n)}$, then a similar coupling argument as in (3.4) can be used. Sample ε_i^* from $\widehat{F}_{(\varepsilon|X_i=x)}(\cdot)$, such that

$$I\!E_{\widehat{F}_{\varepsilon|X=x_i}} \psi(\varepsilon_i^*) = 0 = I\!E_{F_{\varepsilon|X=x_i}} \psi(\varepsilon^\sharp) \tag{3.26}$$

and then

$$|\psi(\varepsilon_i^*) - \psi(\varepsilon_i^\sharp)| = \mathcal{O}_p(h^2\Gamma_n) \tag{3.27}$$

This condition ensures that

$$\frac{n^{-1}\sum_{i=1}^n \{\psi(\varepsilon^*) - \psi(\varepsilon_i^\sharp)\}}{n^{-1}\sum_{i=1}^n \mathbf{f}(X_i \in B_h)} = o_p(h^2\Gamma_n), \tag{3.28}$$

where B_h is a ball of radius h.

This argument is based on two facts. First from (3.26) the means are zero and second that (3.27) holds. The latter can be satisfied only if $\psi(.)$ is bounded.

The above theorem guarantees the closeness of the bootstrap analogue to the estimator.

In the framework of additive model in (3.13), the resampling scheme is considered as in (3.16) and (3.17), a theorem in the same fashion can also be achieved.

Theorem 3.3.2. *Let assumptions A.1- A.7 be fulfilled, then*

$$\sup_{x\in B} |(\widehat{m}_j - m_j)(x) - \{(\widehat{m}_j^* - \widehat{m}_j)(x)\}| = \mathcal{O}_{p(H^2\Gamma_n)}.$$

3.4 Simulation

This section is divided into two parts. First, we concentrate on the univariate $x \in [0, 1]$ case and the bootstrap procedure (3.4), (3.5), check the validity of the bootstrap procedure, and compare it with asymptotic uniform bands developed as in Härdle (1989). Second, we adopt the bootstrap procedure for the additive model as in (3.21), and check the validity of the bootstrap band in the same fashion.

We summarize the bootstrap procedure in the univariate case as following:

- Simulate (X_i, Y_i), $i = 1, \ldots, n$ according to the predefined joint probability density function (pdf) $f(x, y)$. In order to compare with previous literature,

we keep the same setting, the joint pdf of bivariate data,

$$f(x,y) = g\{y-\sin(\pi x)\}\mathbf{f}(x\in[0,1]) \quad (3.29)$$
$$g(u) = 9\varphi(u)/10+\varphi(u/9)/90 \quad (3.30)$$

- Compute the robust smoother $\widehat{l}_h(x)$, $\widehat{\varepsilon}_i \stackrel{\text{def}}{=} Y_i - \widehat{l}_h(X_i)$
- Compute the conditional edf:

$$\widehat{F}_{\varepsilon|X}(t) = \frac{\sum_{i=1}^n K_h(x-X_i)\mathbf{f}(\widehat{\varepsilon}_i\le t)}{\sum_{i=1}^n K_h(x-X_i)}$$

with Gaussian kernel

$$K_h(u) = (\sqrt{2\pi})^{-1}\exp\{-u^2/2h\}/h,$$

$h = 0.06$.

- For each $i=1,\ldots,n$, generate random variable $\varepsilon_i^* \sim \widehat{F}_{(\varepsilon|X)}(t)$, $i=1,\ldots,n$:

$$Y_i^* = \widehat{l}_g(X_i)+\varepsilon_i^*,$$

with $g = 0.2$.

- For each sample $\{X_i, Y_i^*\}_{i=1}^n$, compute $\widehat{l}^*_{h,g}(\cdot)$ and the random variable

$$d_{i^*} \stackrel{\text{def}}{=} \sup_{x\in\Theta}[|\widehat{l}^*_{h,g}(x)-\widehat{l}_g(x)|\sqrt{\widehat{f}_X(x)}I\!E_{Y|X}\{\psi'(\varepsilon_i^*)\}/\sqrt{I\!E_{Y|X}\{\psi^2(\varepsilon_i^*)\}}]$$

- Calculate the $1-\alpha$ quantile d^*_α of $d_1,\ldots,d_{n^*}$.
- Construct the bootstrap uniform band centered around $\widehat{l}_h(x)$

$$\widehat{l}_h(x)\pm[\sqrt{\widehat{f}_X(x)}I\!E_{Y|X}\{\psi'(\varepsilon_i^*)\}/\sqrt{I\!E_{Y|X}\{\psi^2(\varepsilon_i^*)\}}]^{-1}d^*_\alpha$$

Figure 3.1 shows the theoretical signal curve, robust estimation using Huber loss function with corresponding 95% uniform confidence band from the asymptotic theory and the confidence band from the bootstrap. The real curve is marked as the grey solid line. We then compute the classic robust estimate based on asymptotic theory according to Härdle (1989). We notice that the asymptotic band is narrower than the bootstrap band. The width of the bands has not been affected by outliers since we adopt robust estimation with a Tukey biweight loss. To compare which method is more precise, Table 3.1 presents respectively the

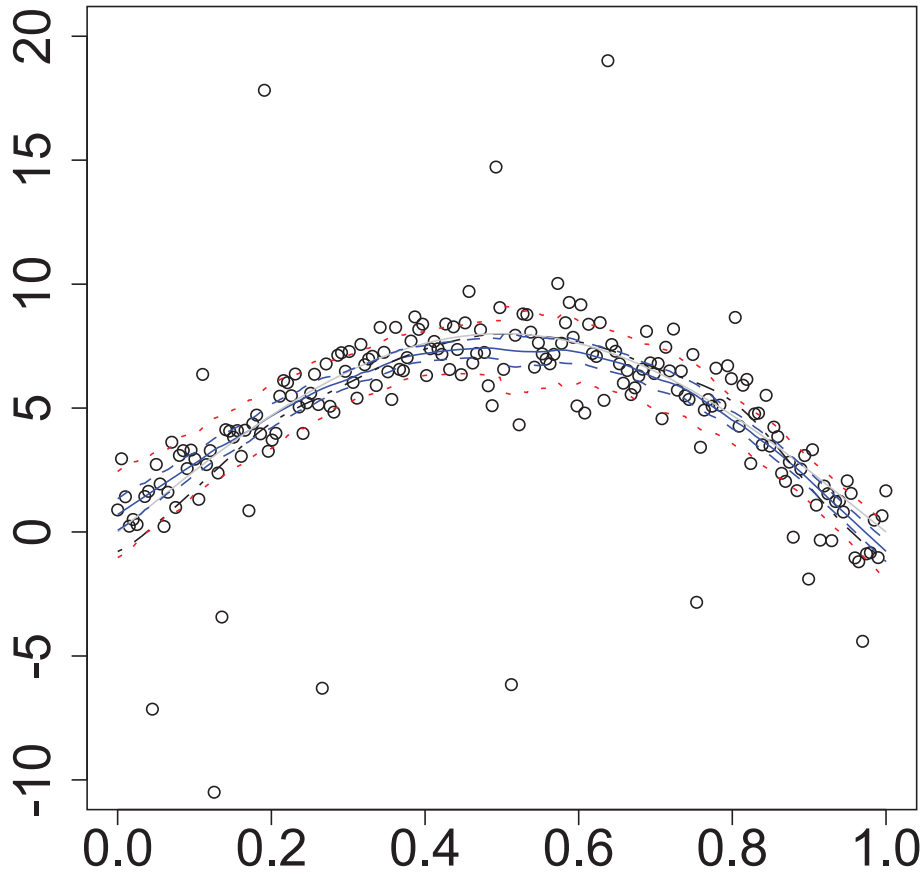


Figure 3.1: Plot of true curve (grey), robust estimation and band (blue dashed), local polynomial estimation (black), bootstrap band (red dotted)

	95%		90%	
n	Cov. Prob.	Area	Cov. Prob.	Area
100	0.88(0.98)	1.23(2.51)	0.83(0.99)	1.02(2.20)
200	0.89(0.98)	0.89(1.95)	0.81(0.96)	0.74(1.76)
400	0.90(0.96)	0.78(1.32)	0.85(0.92)	0.64(1.15)

Table 3.1: Averaged coverage probabilities and areas of nominal asymptotic (bootstrap) with 100 repetitions per sample, and 200 samples.

simulated coverage probabilities together with the calculated area of the 95% and 90% confidence band, for sample size $n = 100, 200, 400$. 100 simulation runs are carried out and 100 bootstrap samples are generated for each simulation. From Table 3.1 , we observe that, for the asymptotic method, coverage probabilities improve a little bit with increased sample size and the bootstrap method (shown in brackets) obtains a larger coverage probability than the asymptotic one, though still a little bit higher than the nominal coverage. It is also observed that the size of the bands decrease with increased sample sizes. Overall, the bootstrap method displays a better convergence rate, while not sacrificing much on the width of the bands.

3.4.1 Additive model

We now extend to multivariate covariates and use an additive model and a different mean function for the estimation. The bootstrap procedure is as follows:

- Simulate (X_i, Y_i), $i = 1, \ldots, n$ following model (3.12) and (3.13). The variable $x_1, x_2, x_3, x_4 \sim U(-2.5, 2.5)$,

  $$m_1(x_1) = \sin(\pi x_1), m_2(x_2) = \Phi(3x_2), m_3(x_3) = x_3^3, m_4(x_4) = x_4^4,$$

  and ε_i is simulated from a mixture normal density function with density $9\varphi(u/10)/100 + \varphi(u)/10$.

- Compute the estimation $\widehat{m}_1(x_1), \widehat{m}_2(x_2), \widehat{m}_3(x_3), \widehat{m}_4(x_4)$ via (3.14) and $\widehat{\varepsilon}_i = Y_i - \sum_{j=1}^4 \widehat{m}_l(x_{i,j})$.

- For each $i = 1, \ldots, n$, generate random variable ε^*_{i,i^*}, $i^* = 1, \ldots, n^*$ as in (3.16), where n^* is the bootstrap sample size:

  $$Y_{i,i^*} = \sum_{j=1}^{4} \widehat{m}_l(x_{i,j}) + \varepsilon^*_{i,i^*}.$$

- For each sample $\{x_{1i}, x_{2i}, x_{3i}, x_{4i}, y_i^*\}$, compute $\widehat{m}_j^*(.)$ and the random variable

  $$d_{i^*} \stackrel{\text{def}}{=} \sup_{x \in [-2.5, 2.5]} \{\sqrt{\widehat{f}_{x_j}(x)} I\!E_{y|x}\{\psi'(\varepsilon_i^*)\} / \sqrt{I\!E_{y|x}\{\psi^2(\varepsilon_i^*)\}} |\widehat{m}_j^*(x) - \widehat{m}_j(x)|\}$$

- Calculate the $1 - \alpha$ quantile d^*_α of $d_1, \ldots, d_{n^*}$.

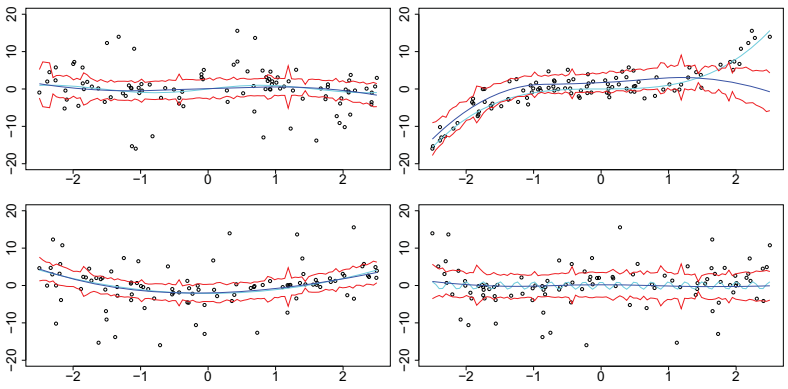


Figure 3.2: Plot of true curve (dark blue), robust estimation and bands (cyran), bootstrap band (red dotted)

- Construct the bootstrap uniform band centered around $\widehat{m}_j(x_j)$

$$\widehat{m}_j(x_j) \pm [\sqrt{\widehat{f}_{x_j}(x_j)}\mathbb{E}_{Y|X_j}\{\psi'(\varepsilon_i^*)\}/\sqrt{\mathbb{E}_{Y|X_j}\{\psi^2(\varepsilon_i^*)\}}]^{-1}d_\alpha^*$$

The estimation of $\widehat{m}_j(x_j)$s $(j = 1, \ldots, 4)$ and their bootstrap confidence bands are shown in Figure 3.2.

The simulated coverage probabilities are shown in Table 3.2. The coverage probabilities are roughly close to the nominal level and the widths of band are clearly shrinking w.r.t. the sample sizes.

3.5 Empirical analysis

3.5.1 Firm expenses analysis

Yafeh & Yosha (2003) use a sample of Japanese firms in the chemical industry to examine whether a concentrated shareholding is associated with lower expenditure on activities with scope for managerial private benefits. In this section, we focus on the same sub regression problem as in Horowitz & Lee (2005). The dependent variable Y is: general sales and administrative expenses deflated by sales (denoted by MH5), which is one of five measures of expenditures on activities with

	n	Cov. Prob.	Area
95%	100	0.95, 0.98, 0.83, 0.95	6.06, 5.37, 5.44, 5.21
	200	0.88, 0.95, 0.93, 0.88	5.50, 4.74, 4.54, 4.65
	400	0.84, 0.95, 0.96, 0.84	4.83, 3.63, 3.76, 3.70
90%	100	0.89, 0.94, 0.85, 0.92	5.88, 5.07, 5.04, 5.30
	200	0.90, 0.95, 0.86, 0.88	4.84, 3.84, 3.85, 4.00
	400	0.85, 0.90, 0.92, 0.84	4.02, 3.25, 3.11, 3.03

Table 3.2: Simulated coverage probabilities and areas of nominal (bootstrap) with 100 repetitions per sample, and 200 samples.

scope for managerial private benefits considered. The covariates are: ownership concentration (denoted by TOPTEN, cumulative shareholding by the largest ten shareholders), and firm characteristics: the log of assets, firm age, and leverage (the ratio of debt to debt plus equity), sample size= 185. The regression model we consider here is:

$$\begin{aligned} MH5 &= m_0 + m_1(TOPTEN) + m_2\{log(Assets)\} \\ &\quad + m_3(Age) + m_4(Leverage) + error \end{aligned}$$

The estimated additive components and its bootstrap confidence bands are shown in Figure 3.3. Similarly, it can be seen that the nonlinear effects are log(asset) and TOPTEN, and the firm age effects are minor compared to the other three. Differently, the effect of leverage is also a little bit nonlinear, and the shape of curves deviates from what Horowitz & Lee (2005) present, especially for the effect of TOPTEN. This may due to the different subjects studied: in our case robust estimation with Tukey biweight loss, while in their case the conditional median curve.

3.5.2 The impact on stock market

We analyze how the four markets (oil, currency, bond, real estate) affects the stock market. This study would give implications to the interactions of the economic conditions from different sectors. The data source is ProQuest Statistical Datasets (http://www.lnstatistical.com
/Main.jsp;jsessionid=009E36E74DFA15C80B74EE0BDAEB5746), we focus on the US market. Therefore, the covariates are taken as: the crude oil price, EUR- USD exchange rate, the 10 year treasury constant maturity inflation index %, the real estate price, and the Y variable is S&P 500 index returns. The data are synchronized to weekly frequency. We select the data during the period $20080903 - 20111128$.

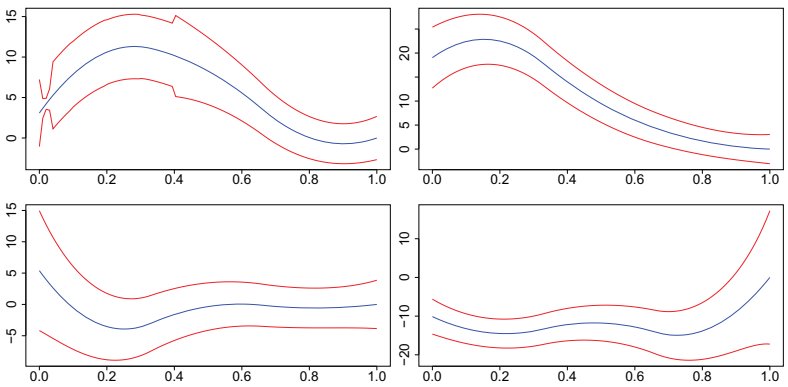


Figure 3.3: Robust estimation (blue), bootstrap band (red dotted), left up: Log(Asset), right up: Leverage, left below: Age, right below: TOPTEN.

It can be observed that all the four markets have non linear effects on the stock indices values, Figure 3.4, but only exchange rate EUR-USD and crude oil prices affect the the stock indices returns nonlinearly, Figure 3.5. It is not difficult to interpret the relationships: In Figure 3.4, for the exchange rate EUR-USD, the weakness of EUR up to a certain level (< 1.27) are negatively correlated with the stock indices, and then a positive correlation follows, but this relationship is again reversed when the EUR is too high(> 1.43). Oil prices have negative impact on stock indices at every level, but the effects decrease when the prices raise. As for the inflation index, when the inflation rate is high, interest rates are typically high, this may reduce the consumption and investments in the stock market. So one sees a negative correlation there when the inflation index is bigger than (0.7). Finally, increasing real estate prices can be a sign of booming economic condition, therefore the stock indices raise when the real estate prices get higher. However, when the real estate prices are too high, it is likely that there exist bubble, so one sees a drop in the market indices.

In Figure 3.5, we see difference effects on S&P log returns, exchange rate EUR-USD are positively correlated with returns until a high lever (> 1.40), the crude oil has majorally negative effects on stock returns. More nonlinearity is presented in the plots for inflation index and real estate price.

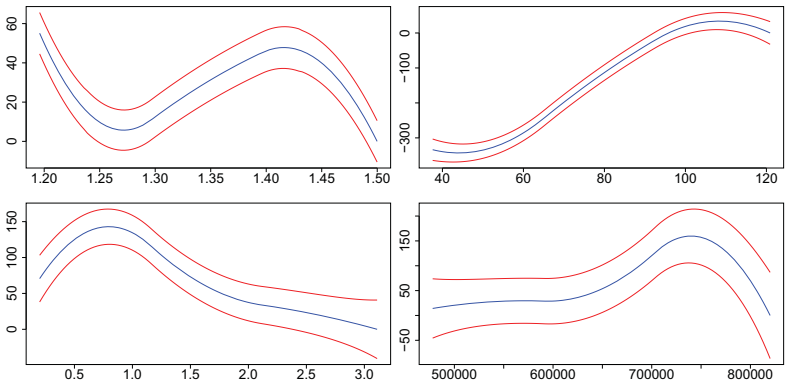


Figure 3.4: Robust estimation (blue), bootstrap band (red dotted), Y: S&P index, left up: exchange rates EUR-USD, right up: crude oil price, left below: inflation index, right below: real estate price.

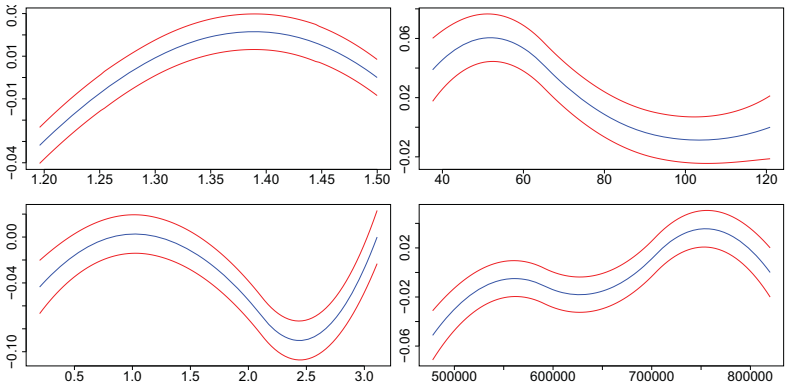


Figure 3.5: Robust estimation (blue), bootstrap band (red dotted), Y: S&P index log return, left up: exchange rates EUR-USD, right up: crude oil price, left below: inflation index, right below: real estate price.

3.6 Conclusion

We have developed and proved the bootstrap improvement for a wide class of smoothers with bounded influence function. Moreover, we extend our results to additive models to cope with curse of dimensionality.

3.7 Appendix

3.7.1 Proof of Theorem 3.1

To prove Theorem 3.1, we need to show:

$$\max_i[(\widehat{l}_h - l)(X_i) - \{(\widehat{l}^*_{h,g} - \widehat{l}_g)(X_i)\}] = \mathcal{O}_{p(h^2\Gamma_n)} \tag{3.31}$$

Proving (3.31) can be done by showing the following conditions:

$$\max_i |\varepsilon_i^\sharp - \varepsilon_i^*| = \mathcal{O}_p(h^2\Gamma_n) \tag{3.32}$$

$$\max_i |\psi(\varepsilon^*) - \psi(\varepsilon^\sharp)| = \mathcal{O}_p(h^2\Gamma_n) \tag{3.33}$$

Let us start with (3.32), set

$$r = \widehat{F}^{-1}_{(\varepsilon_i|X_i)}(t) - F^{-1}_{(\varepsilon_i|X_i)}(t)$$

Since $f_{\varepsilon_i|X_i}(\cdot)$ is continuous and bounded from below in $[-b, b]$(A.2), we have

$$\widehat{F}_{\varepsilon|X}(t) = F_{(\varepsilon|X)}(t - r) \leq F_{(\varepsilon|X)}(t) - c_1 r, \quad \forall t \in [-b, b]$$

for some constant c_1.

From Franke & Mwita (2011), we have, with assumption A.1-A.4, for any small enough (positive) $b \to 0$,

$$\sup_{|t|\leq b, i=1,\ldots,n} |\widehat{F}_{(\varepsilon|X)}(t) - F_{(\varepsilon|X)}(t)| = \mathcal{O}_p(h^2\Gamma_n b^{1/2} + b^2), \tag{3.34}$$

Hence,

$$h^2\Gamma_n \geq \sup_{|t|\leq b} |\widehat{F}_{(\varepsilon|X)}(t - r) - F_{(\varepsilon|X)}(t)| \geq c_1 r,$$

which further implies that

$$\sup_{u\in B} |\widehat{F}^{-1}_{(\varepsilon|X)}(t) - F^{-1}_{(\varepsilon|X)}(t)| \leq h^2\Gamma_n, \forall u \in [0, 1].$$

Therefore, (3.32) is proved, and (3.33) is a direct consequence of (3.32), since $\psi(.)$ is Lipchitz continuous (A.1),

$$|\psi(\varepsilon^*) - \psi(\varepsilon^\sharp)| \leq C|\varepsilon_i^* - \varepsilon_i^\sharp|.$$

Define now

$$\begin{aligned}
G_n^*(\theta, X_i) &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n W_{h,j}(X_i)[\psi\{\varepsilon_j^* - \theta + \widehat{l}_g(X_j)\}] \\
&= \frac{1}{n}\sum_{j=1}^n W_{h,j}(X_i)\{\psi(Y_j^* - \theta)\} \\
G_n^\sharp(\theta, X_i) &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^n W_{h,j}(X_i)[\psi\{\varepsilon_j^\sharp - \theta + l(X_j)\}] \\
&= \frac{1}{n}\sum_{j=1}^n W_{h,j}(X_i)\{\psi(Y_j^\sharp - \theta)\} \\
T_n(X_i) &\stackrel{\text{def}}{=} G_n^*\{\widehat{l}_g(X_i), X_i\} - G_n^\sharp\{l(X_i), X_i\}.
\end{aligned}$$

Note that,

$$I\!E_{\widehat{F}_{\varepsilon|X_i=x}}\psi(\varepsilon_i^*) = 0 = I\!E_{F_{\varepsilon|X_i=x}}\psi(\varepsilon^\sharp). \tag{3.35}$$

According to A.1, the Lipschitz condition of the function $\psi(\cdot)$,

$$\begin{aligned}
&\max_i |T_n(X_i)| \\
= &\max_i |\frac{1}{n}\sum_{j=1}^n W_{h,j}(X_i)\{\psi(\varepsilon_j^* - \widehat{l}_g(X_i) + \widehat{l}_g(X_j)) - \psi(\varepsilon_i^\sharp - l(X_i) + l(X_j))\}| \\
\leq &\max_i |\frac{1}{n}\sum_{j=1}^n W_{h,j}(X_i)(C[\{\varepsilon_j^* - \widehat{l}_g(X_i) + \widehat{l}_g(X_j)\} - \{\varepsilon_i^\sharp - l(X_i) + l(X_j)\}])|
\end{aligned}$$

So we can break the upper bound by two terms

$$\begin{aligned}
\max_i |T_n(X_i)| \leq &\max_i |\frac{1}{n}C\sum_{j=1}^n W_{h,j}(X_i)(\varepsilon_j^* - \varepsilon_i^\sharp)| \\
+ &\max_i |\frac{1}{n}C\sum_{j=1}^n W_{h,j}(X_i)[\{\widehat{l}_g(X_j) - \widehat{l}_g(X_j)\} - \{l(X_j) + l(X_i)\}]| \\
\leq &\max_i T_{n,1}(X_i) + \max_i T_{n,2}(X_i)
\end{aligned}$$

$\max_i |T_{n,1}(X_i)|$ is know to have the rate $\mathcal{O}_p(n^{-1}h^{3/2}\Gamma_n)$, and

$$\max_i T_{n,2}(X_i) \leq \max_i |\frac{1}{n} C \sum_{j=1}^{n} W_{h,j}(X_i)\{\widehat{l}'_g(X_{i,j,0}) - \widehat{l}'(X_{i,j,0})\}(X_i - X_j)|,$$

where $X_{i,j,0}$ is a point between X_i and X_j, and C is a constant.

$\sup_{x\in B} \|\widehat{l}_g(x) - \widehat{l}(x)\|$ is typically of the rate $\mathcal{O}_p(g^{-1}(ng)^{-1/2}\Gamma_n + g^3)$, see Stone (1982). Therefore the optimal rate for g would be $\mathcal{O}(n^{-1/9})$ in our case (as in A.4). Then we can achieve

$$\max_i T_{n,2}(X_i) = o_p(h^2\Gamma_n)$$

We know that,

$$\widehat{l}^*_{h,g}(X_i) - \widehat{l}_g(X_i) = -\frac{G^*_n\{\widehat{l}_g(X_i), X_i\}}{G'^*_n\{\widehat{l}_g(X_i), X_i\}} + o_{p(h^2),} \tag{3.36}$$

$$\widehat{l}_h(X_i) - l(X_i) = -\frac{G^\sharp_n\{l(X_i), X_i\}}{G'^\sharp_n\{l(X_i), X_i\}} + o_{p(h^2).} \tag{3.37}$$

This means,

$$\begin{aligned}
&|\widehat{l}^*_{h,g}(X_i) - \widehat{l}_g(X_i) - \widehat{l}_h(X_i) + l(X_i)| \\
&= -\frac{G^*_n\{\widehat{l}_g(X_i), X_i\}}{G'^*_n\{\widehat{l}_g(X_i), X_i\}} + \frac{G^\sharp_n\{l(X_i), X_i\}}{G'^\sharp_n\{l(X_i), X_i\}} + o_{p(h^2)} \\
&= -\frac{G'^\sharp_n\{l(X_i), X_i\}[G^*_n\{\widehat{l}_g(X_i), X_i\} - G^\sharp_n\{l(X_i), X_i\}]}{G'^*_n\{\widehat{l}_g(X_i), X_i\}G'^\sharp_n\{l(X_i), X_i\}} \\
&-\frac{G^\sharp_n\{l(X_i), X_i\}[G'^*_n\{\widehat{l}_g(X_i), X_i\} - G'^\sharp_n\{l(X_i), X_i\}]}{G'^*_n\{\widehat{l}_g(X_i), X_i\}G'^\sharp_n\{l(X_i), X_i\}} + o_{p(h^2)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\max_i |\widehat{l}^*_{h,g}(X_i) - \widehat{l}_g(X_i) - \widehat{l}_h(X_i) + l(X_i)| \\
&= \mathcal{O}(\max_i T_n(X_i)) + o_p(h^2\Gamma_n) + o_{p(h^2\Gamma_n),}
\end{aligned}$$

and (3.31) is proved.

The claim (3.24) can be proved from (3.31) using the fact that,

$$\sup|A_n(x)| \leq \max_i|A_n(X_i)| + \max_i \sup_{x\in[X_i,X_{i+1}]}|A_n(X_i) - A(x)| \tag{3.38}$$

it suffices to consider the speed of the last term.

With Lipschitz continuity of $A_n(.)$:

$$\max_i \sup_{x\in[X_i,X_{i+1}]}|A_n(X_i) - A(x)| \leq c_2 \max_i \sup_x |X_i - x|, \tag{3.39}$$

where $c_2 > 0$ is a constant, this upper random bound is of order $\mathcal{O}_p(n^{-1/d}\log n) = \mathcal{O}_{p(h^2\Gamma_n)}$

The uniform bound for $\|X_i - x\|$ results from the uniform law of large numbers over a ball of size $n^{-1/d}$, see Penrose (1964), Theorem 1.1.

Therefore, Theorem 3.1 is proved.

3.7.2 Proof of Theorem 3.2

The number of regressors in (3.13) is of order $p = dL + 1$ (more precisely $p = \sum_{j=1}^{d} L_j$) with $L \sim n^{1/5}$. To simply our setting, assume $L_j = L$.

Portnoy (1997) shows that as long as $n^{-1}(p\log n)^{3/2} \to 0$ then the estimators of the regression parameters are consistent and have the standard variance. In our situation,

$$n^{-1}n^{1/5*2/3}\log n = \mathcal{O}_{(1)} \tag{3.40}$$

and therefore the condition is satisfied.

Now we have a look at the behavior of the design matrix in (3.14).

$$\begin{aligned}
\widehat{\mathcal{L}}(A) &\stackrel{\text{def}}{=} -n^{-1}\sum_{i=1}^{n}\rho\{Y_i - A^\top \Phi(X_i)\} \\
\nabla\widehat{\mathcal{L}}(A) &\stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^{n}\psi\{Y_i - A^\top \Phi(X_i)\}\Phi(X_i) \\
\nabla^2\widehat{\mathcal{L}}(A) &\stackrel{\text{def}}{=} -n^{-1}\sum_{i=1}^{n}\rho''\{Y_i - A^\top \Phi(X_i)\}\Phi(X_i)\Phi(X_i)^\top
\end{aligned}$$

Recall that

$$\widehat{A} = \text{argmin}_A \widehat{\mathcal{L}}(A)$$

Lemma 14 of Stone (1985) ensures that with probability approaching 1, $\widehat{A}$ exists uniquely and that $\nabla\widehat{\mathcal{L}}(\widehat{A}) = 0$.

In addition, there exists $\overline{m}(x) = \overline{A}^\top \Phi(x)$, such that

$$\overline{m}(x) = \overline{A}^\top \Phi(x)$$

and

$$\sup_{x\in B}|\overline{m}(x)-m(x)|\leq C_{\infty}H^{2} \tag{3.41}$$

According to the bounded influence condition A.1,

$$\begin{aligned}\nabla\widehat{\mathcal{L}}(\overline{A}) &= -n^{-1}\sum_{i=1}^{n}\psi\{-\overline{A}^{\top}\Phi(X_i)+\varepsilon_i+A^{\top}\Phi(X_i)\}\Phi(X_i)\\ &= n^{-1}\sum_{i=1}^{n}[\psi(0)+\psi'(0)\{A^{\top}\Phi(X_i)+\varepsilon_i-\overline{A}\Phi(X_i)\}]\Phi(X_i)+\mathcal{O}_{a.s.}(H^{3})\\ &\leq n^{-1}\sum_{i=1}^{n}M\{A^{\top}\Phi(X_i)+\varepsilon_i-\overline{A}\Phi(X_i)\}\Phi(X_i)+\mathcal{O}_{a.s.}(H^{3})\end{aligned}$$

We know first that,

$$I\!E|\{m(X_i)-\overline{m}(X_i)\}\Phi(X_i)|=\mathcal{O}(H^{2}).$$

Let $\xi_{i,j}\stackrel{\text{def}}{=}|m(X_i)-\overline{m}(X_i)||\Phi(X_i)|-I\!E|m(X_i)-\overline{m}(X_i)||\Phi(X_i)|$.
By Bernstein's Lemma:

Lemma 3.7.1. *Let $Z_1,\ldots,Z_n$ be independent r.v.s.*

$$\log I\!E\exp(tZ_i)\leq I\!E(Z_i^{2})t^{2}/2$$

for all $t\in[0,\infty]$. Then

$$I\!P\Big(|\sum_{i=1}^{n}X_i|\geq t\sqrt{2\sum_{i=1}^{n}I\!EX_i^{2}}\Big)\leq 2\exp(-t^{2})$$

Therefore, we can derive that,

$$n^{-1}\sum_{i=1}^{n}\xi_{i,j}=\mathcal{O}_{a.s.}(H^{2}n^{-1/2}\Gamma_n) \tag{3.42}$$

The last term

$$n^{-1}|\sum_{i=1}^{n}\varepsilon_i\Phi(X_i)|=\mathcal{O}_{a.s.}(n^{-1/2}\Gamma_n) \tag{3.43}$$

Therefore, one has collective term from (4.33) and (3.43),

$$\|\nabla\widehat{\mathcal{L}}(\overline{A})\|=\mathcal{O}_{a.s.}(H^{3/2}+H^{-1/2}n^{-1/2}\Gamma_n),$$

where $\|.\|$ denotes the Euclidean norm.

By assumption A.5, A.7, $\forall l = 1, \ldots, L$, the d dimensional vector $\Phi_l^\top(X_i)$ satisfies,

$$\beta\|b\|^2/d \geq I\!E b^\top \Phi_l^\top(X_i)\Phi_l(X_i)b \geq \alpha\|b\|^2/d,$$

where α and β are two constants.

Lemma 3.7.2. *Assume A.1 and A.7, as $n \to \infty$,*

$$\begin{aligned} \|\widehat{A} - A\| &= \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\Gamma_n) \\ max_{i\in 1,\ldots,n}\|\widehat{m}(X_i) - \overline{m}(X_i)\| &= \mathcal{O}_{a.s.}(H + H^{-1}n^{-1/2}\Gamma_n) \end{aligned}$$

Proof According to the mean value theorem, exists an $N_d \times N_d (N_d \stackrel{\text{def}}{=} (L+1)d + 1)$, $\widehat{A}_0 = t\widehat{A} + (1-t)\overline{A}$,

$$\nabla\widehat{\mathcal{L}}(A) - \nabla\mathcal{L}(\overline{A}) = \nabla^2\widehat{\mathcal{L}}(\widehat{A}_0)(\widehat{A} - A),$$

which will lead to

$$\|\widehat{A} - \overline{A}\| = \|\{\nabla^2\widehat{\mathcal{L}}(\widehat{A}_0)\}^{-1}\{-\nabla\widehat{\mathcal{L}}(\overline{A})\}\|,$$

since

$$\nabla^2\widehat{\mathcal{L}}(X_i) = n^{-1}\sum_{i=1}^{n}\Phi(X_i)\Phi(X_i)^\top\rho''(Y_i - \widehat{A}_0^\top\Phi(X_i))$$

According to assumption A.7,

$$c_3 I \geq \nabla^2\widehat{\mathcal{L}}(X_i) \geq c_4 I$$

therefore

$$\|\widehat{A} - \overline{A}\| = \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\Gamma_n).$$

Moreover, by Cauchy-Schwarz inequality

$$\begin{aligned} \max_{i\in 1,\ldots,n}|\widehat{m}(X_i) - \overline{m}(X_i)| &\leq \|\widehat{A} - \overline{A}\|\sup_{x\in[0,1]^d}\|\Phi(x)\| \\ &= \mathcal{O}_{a.s.}(H^{3/2} + H^{-1/2}n^{-1/2}\Gamma_n)\mathcal{O}_{a.s.}(H^{-1/2}) \\ &= \mathcal{O}_{a.s.}(H + H^{-1}n^{-1/2}\Gamma_n). \end{aligned}$$

We would like to check for the pseudo observations $Y_i^\sharp = \overline{m}(X_i) + \varepsilon_i^\sharp$.

$$\begin{aligned} &\sup_{x\in B}|(\widehat{m}_k^*(x) - \widehat{m}_(x)) - (\widehat{m}_k(x) - \overline{m}_k(x))| \\ &\leq |\widehat{A}^* - \widehat{A} - \widehat{A} + \overline{A}|\sup_{x\in[0,1]}|\Phi(x)| \\ &\leq |\nabla^2\widehat{\mathcal{L}}^*(\widehat{A}_0^*)\{-\nabla\widehat{\mathcal{L}}^*(\widehat{A})\} - \nabla^2\widehat{\mathcal{L}}(\widehat{A}_o)\{-\nabla\widehat{\mathcal{L}}(\overline{A})\}|\sup_{x\in[0,1]}|\Phi(x)| \end{aligned}$$

as $\|\nabla^2\widehat{\mathcal{L}}(\widehat{A}_o)\|$ and $\|\nabla^2\widehat{\mathcal{L}}^*(\widehat{A}_0^*)\|$ are both bounded,

$$
\begin{aligned}
&\sup_{x\in B}|(\widehat{m}_k^*(x)-\widehat{m}_(x))-(\widehat{m}_k(x)-\overline{m}_k(x))|\sup_{x\in[0,1]}|\Phi(x)|\\
&\leq C\|n^{-1}\sum_{i=1}^{n}\{\psi(\varepsilon_i^*)-\psi(\varepsilon_i^\sharp)\}\Phi(X_i)\|\sup_{x\in[0,1]}|\Phi(x)|\\
&\leq C\|n^{-1}\sum_{i=1}^{n}\{\psi(\varepsilon_i^*)-\psi(\varepsilon_i^\sharp)\}\Phi(X_i)\|\sup_{x\in[0,1]}|\Phi(x)|\\
&\leq CM\|n^{-1}\sum_{i=1}^{n}\{\varepsilon_i^*-\varepsilon_i^\sharp\}\Phi(X_i)\|\sup_{x\in[0,1]}|\Phi(x)|\\
&\leq CM\|n^{-1}\sum_{i=1}^{n}\{Z_i\widehat{\varepsilon}_i-Z_i\eta_i\}\Phi(X_i)\|\sup_{x\in[0,1]}|\Phi(x)|\\
&=\mathcal{O}_{a.s.}(n^{-1/2}(H+H^{-1}n^{-1/2}\Gamma_n)H^{-1/2})\\
&=o_{a.s.}(H^2\Gamma_n)
\end{aligned}
$$

Chapter 4

Hidden Markov structures for dynamic copulae

4.1 Introduction

Modelling high-dimensional time series is an often underestimated exercise of routine econometrical and statistical work. This slightly pejorative attitude towards day to day statistical analysis is unjustified since actually the calibration of time series models in high dimensions for standard data sizes is not only difficult on the numerical side but also on the mathematical side. Computationally speaking, integrated models for high dimensional time series become more involved when the parameter space is too large. An example is the multivariate GARCH(1,1) BEKK model that for even two dimensions has an associated parameter space of dimension 12. For moderate sample sizes, the parameter space dimension might well be in the range of the sample size or even bigger. This data situation has evoked a new strand of literature on dimension reduction via penalty methods.

In this chapter we take a different route, by calibrating an integrated dynamic model with unknown dependency structure among the d dimensional time series variables. More precisely, the unknown dependency structure may vary within a set of given dependencies. The specific dependence at each time t is unknown to the data analyst, but depends on the dependency pattern at time $t-1$. Therefore, hidden Markov models (HMM) naturally come into play. This leaves us with the problem of specifying the set of dependencies.

An approach based on assuming a multivariate Gaussian or mixed normal is handicapped in capturing important types of data features such as heavy tails, asymmetry, and nonlinear dependencies. Such a simplification might in practice be too restrictive an assumption and might lead to biased results. Copulae are one pos-

sible approach to solving these problems, see Joe (1996). Moreover, copulae allow us to separate the marginal distributions and the dependency model, see Sklar (1959). In recent decades, copula-based models have gained popularity in various fields like finance, insurance, biology, hydrology, etc. Nevertheless, many basic multivariate copulae are still too restrictive and a simple extension by putting in more parameters would lead to the extreme of a totally nonparametric approach that runs into the problem of the curse of dimensionality. A natural compromise is the class of hierarchical Archimedean copulae (HAC). An HAC allows a rich copula structure with a finite number of parameters. Recent works which have shown their flexibility are McNeil & Nešlehová (2009), Okhrin, Okhrin & Schmid (2009), Whelan (2004).

Many attempts have been made to obtain insights into the dynamics of the copulae: Chen & Fan (2005) assumes the underlying sequence is Markovian; Patton (2004) considers an asset-allocation problem with a time-varying parameter of bivariate copulae; Rodriguez (2007) studies financial contagion using switching-parameter bivariate copulae. A likelihood based local adaptive method is an alternative approach for understanding the time evolution, see Giacomini, Härdle & Spokoiny (2009), Härdle, Okhrin & Okhrin (2012). Figure 4.1 presents the LCP (local change point method) window analysis of HAC for exchange rate data. One observes that the structure (upper panel) very often remains the same for a long time, and the parameters (lower panel) are only slowly varying over time. This indicates that the dynamics of HAC functions is likely to be driven by a Markovian sequence connected with the structures and parameter values. This suggests to us a different path of modeling the dynamics: instead of taking a local point of view, we adopt a global dynamic model HMM for the change of both the tree structure and the parameters of the HAC along the time horizon. In this situation, a stochastic process Y with a not directly observable underlying Markov process X is needed to determine the state of distributions of Y. This has been widely applied to speech recognition, see Rabiner (1989), molecular biology, and digital communications over unknown channels. For estimation and inference issues in HMM, see Bickel, Ritov & Rydén (1998) and Fuh (2003), among others.

In this chapter, we propose a new type of dynamic model, called HMM HAC, by incorporating HAC into an HMM framework. The theoretical problems such as parameter consistency and structure consistency are solved. The expectation maximization (EM) algorithm is developed in this framework for parameter estimation. See Section 2 for the model description, Section 3 for theorems about consistency. EM algorithm and computation issues are in Section 4. Section 5 is for the simulation study, and Section 6 is for applications. The technical details are put into the Appendix.

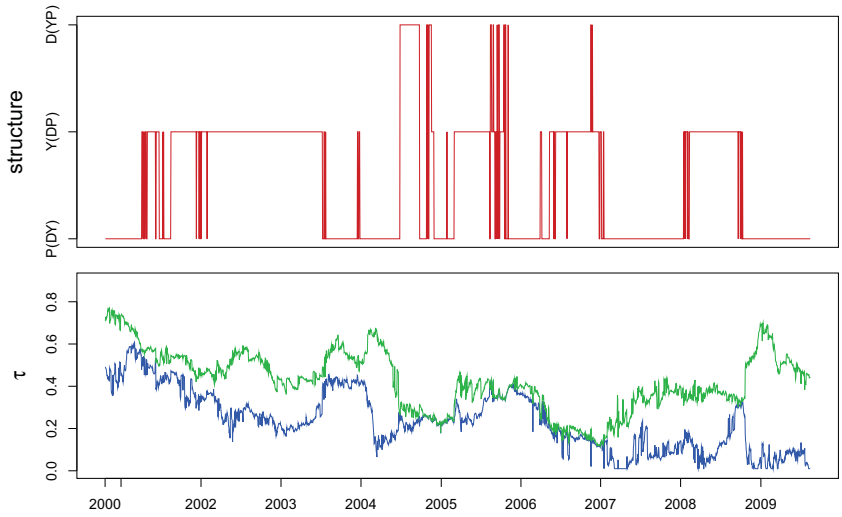


Figure 4.1: LCP for exchange rates: structure (upper) and parameters (lower, θ_1(green) and θ_2)(blue) for Gumbel HAC. $m_0 = 40$.

4.2 Model Description

4.2.1 Incorporating HAC into HMM

A hidden Markov model is a parameterized Markov random walk with an underlying Markov chain viewed as missing data, as in Leroux (1992), Bickel et al. (1998), and Gao & Song (2011). Specifically, in our HMM HAC framework, let $\{X_t, t \geq 0\}$ be a stationary Markov chain on a finite state space $D = \{1, 2, \ldots, M\}$, with transition probability matrix $P = \{p_{ij}\}_{i,j=1,\ldots,M}$ and initial distribution $\pi = \{\pi_i\}_{i=1,\ldots,M}$.

$$\begin{aligned} I\!P(X_0 = i) &= \pi_i, && (4.1)\\ I\!P(X_t = j | X_{t-1} = i) &= p_{ij} && (4.2)\\ &= I\!P(X_t = j | X_{t-1} = i, X_{t-2} = x_{t-2}, \ldots, X_1 = x_1, X_0 = x_0), \\ &\quad i, j = 1, \ldots, M \end{aligned}$$

Let $\{Y_t, t \geq 0\}$ be the associated observations, and they are adjoined with $\{X_t, t \geq 0\}$ in such a way that given $X_t = i, i, \ldots, M$, the distribution of Y_t is fixed:

$$\begin{aligned} I\!P(X_t | X_{1:(t-1)}, Y_{1:(t-1)}) &= I\!P(X_t | X_{t-1}) && (4.3)\\ I\!P(Y_t | Y_{1:(t-1)}, X_{(1:t)}) &= I\!P(Y_t | X_t), && (4.4) \end{aligned}$$

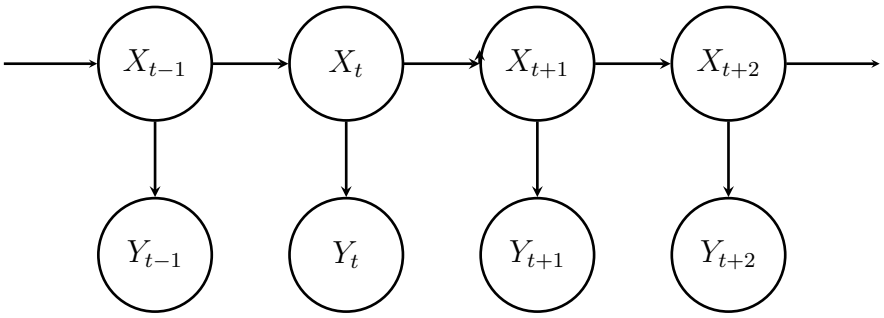


Figure 1: Graphical representation of the dependence structure of HMM
Figure 4.2: Graphical representation of the dependence structure of HMM, where X_t depends only on X_{t-1} and Y_t only on X_t.

where $Y_{1:(t-1)}$ stands for $\{Y_1, \ldots, Y_{t-1}\}$, $t < T$.

Let $f_j\{\cdot; \boldsymbol{\theta}^{(j),s^{(j)}}\}$ be the conditional density of Y_t given X_{t-1}, $X_t = j$ with $\boldsymbol{\theta} \in \Theta, \mathbf{s} \in S$, $j = 1, \ldots, M$ being the unknown parameters. That is, $\{X_t, t \geq 0\}$ is a Markov chain, given $X_0, X_1, \ldots, X_T$, with $Y_0, Y_1, \ldots, Y_T$ being independent. Note that $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}) \in \mathbb{R}^{dM}$ are the unknown dependency parameters, $\mathbf{s} = (s^{(1)}, \ldots, s^{(M)})$ are the unknown HAC structure parameters, and its true value is denoted by $\boldsymbol{\theta}^*$ and $\mathbf{s}^*$. See Figure 4.2 for a graphical illustration, and in Appendix 7.2 we have a more strict formulation of the definition of a HMM.

For given d dimensional time series $y_1, \ldots, y_T, \in \mathbb{R}^d$ ($y_t = (y_{1t}, y_{2t}, y_{3t}, \ldots, y_{dt})^\top$) connected with unobservable (or missing) $x_1, \ldots, x_T$ from the given hidden Markov model, define π_{x_t} as the π_i for $x_0 = i, i = 1, \ldots, M$, and $p_{x_{t-1}x_t} = p_{ji}$ for $x_{t-1} = j$ and $x_t = i$. The full likelihood function given one realization of $\{x_t, y_t\}_{t=1}^T$ is

$$p_T(y_1, \cdots, y_T; x_1, \ldots, x_T) = \pi_{x_0} \prod_{t=1}^{T} p_{x_{t-1}x_t} f_{x_t}(y_t; \boldsymbol{\theta}^{(x_t)}, s^{(x_t)}), \tag{4.5}$$

and the likelihood for only the observations $\{y_t\}_{t=1}^T$ by marginalization:

$$p_T(y_{1:T}) = \sum_{x_0=1}^{M} \cdots \sum_{x_n=1}^{M} \pi_{x_0} \prod_{t=1}^{T} p_{x_{t-1}x_t} f_{x_t}(y_t; \boldsymbol{\theta}^{(x_t)}, s^{(x_t)}), \tag{4.6}$$

with the abbreviation of $p_T(y_1, \cdots, y_T)$ as $p_T(y_{1:T})$.

The novelty of our approach lies in a special parametrization of $f_{x_t}(y_t; \boldsymbol{\theta}^{(x_t)}, s^{(x_t)})(x_t = i)$ (abbreviated as $f_i(.)$), which helps to properly understand the dynamics of a multivariate distribution. Up to now, typical parameterizations have been mixtures of log-concave or elliptical symmetric densities, such as those from Gamma

or Poisson families, which are not flexible enough to model high dimensional time series. The advantage of the copula is that it splits the multivariate distribution into its margins and a pure dependency component. In other words, it captures the dependency between variables eliminating the impact of the marginal distributions. Technical details and properties about copulae are to be found in the Appendix 4.7.1.

Furthermore, we incorporate this procedure into the HMM framework. We denote the underlying Markov variable X_t as a dependency type variable. If $x_t = i$, the parameters $(\boldsymbol{\theta}^{(i)}, s^{(i)})$ determined by state $i = 1, \ldots, M$ take values on $S \times \Theta$, where S is a set of discrete candidate states corresponding to different dependency structures of the HAC, and Θ is a compact set in $\mathbb{R}^{d-1}$ wherein the HAC parameters take their values. Therefore,

$$f_i(\cdot) = c\{F_1^{\mathfrak{m}}(y_1), F_2^{\mathfrak{m}}(y_2), \ldots, F_d^{\mathfrak{m}}(y_d), \boldsymbol{\theta}^{(i)}, s^{(i)}\} f_1^{\mathfrak{m}}(y_1) f_2^{\mathfrak{m}}(y_2) \cdots f_d^{\mathfrak{m}}(y_d), \tag{4.7}$$

with $f_i^{\mathfrak{m}}(y_i)$ the marginal densities, $F_i^{\mathfrak{m}}(y_i)$ the marginal cdf, $c(\cdot)$ the copula density, and see more details in Appendix 7.1.

Let $\boldsymbol{\theta}^{(i)} = (\theta_{i1}, \ldots, \theta_{i,d-1})^\top$ be the dependency parameters of the copulae starting from the lowest up to the highest level connected with a fixed state $x_t = i$ and the $f_i(.)$. The multistage maximum likelihood estimator $(\widehat{\boldsymbol{\theta}}^{(i)}, \widehat{s}^{(i)})$ solves the system

$$\left(\frac{\partial \mathcal{L}_1}{\partial \theta_{i1}}, \ldots, \frac{\partial \mathcal{L}_{d-1}}{\partial \theta_{id-1}} \right)^\top = \mathbf{0}, \tag{4.8}$$

$$\begin{aligned} \text{where} \quad \mathcal{L}_j &= \sum_{t=1}^{T} w_{it} l_{ij}(Y_t), \text{ for } j = 1, \ldots, d-1, \\ l_{ij}(Y_t) &= \log \Big(c\big[\{\widehat{F}_m^{\mathfrak{m}}(y_{tm}, \boldsymbol{\alpha}_m)\}_{m \in \{1,\ldots,d\}};\, s^{(j)}, \{\theta_{i\ell}\}_{\ell=1,\ldots,d-1}\big] \\ &\qquad \prod_{m \in \{1,\ldots,d\}} \widehat{f}_m^{\mathfrak{m}}(y_{tm}, \boldsymbol{\alpha}_m) \Big) \\ &\qquad \text{for } j = 1, \ldots, d-1,\ t = 1, \ldots, T. \end{aligned} \tag{4.9}$$

where $\widehat{F}_m^{\mathfrak{m}}(\cdot)$ is an estimator (either nonparametric or parametric) of the marginal cdf $F_m^{\mathfrak{m}}(\cdot)$ and if the estimated margins are parametrical, then $\widehat{F}_m^{\mathfrak{m}}(\cdot) = F_m^{\mathfrak{m}}(\cdot, \widehat{\boldsymbol{\alpha}}_m)$. The marginal densities $\widehat{f}_m^{\mathfrak{m}}(\cdot)$ are estimated according to the cdfs, and w_{it} is the weight associated with state i and time t, see (4.15). Chen & Fan (2006) and Okhrin et al. (2009) provide the asymptotic behavior of the estimates.

4.2.2 Likelihood estimation

For the estimation of the HMM HAC model, we adopt the EM algorithm, Dempster, Laird & Rubin (1997). In the context of HMM, the EM algorithm is also known as the Baum–Welch algorithm. Let us recall the description in the setting of HMM on HAC.

Recall the full likelihood $p_T(y_{1:T}; x_{1:T})$ in (4.5) and the partial likelihood $p_T(y_1, \ldots, y_T)$ in (4.6), and the log likelihood:

$$\log\{p_T(y_1, \ldots, y_T)\} = \log\{\sum_{x_0=1}^{M} \cdots \sum_{x_n=1}^{M} \pi_{x_0} \prod_{t=1}^{T} p_{x_{t-1}x_t} f_{x_t}(y_t; \boldsymbol{\theta}^{(x_t)}, s^{(x_t)})\} \qquad (4.10)$$

The EM algorithm suggests estimating a sequence of parameters $\mathfrak{g}_{(i)} \stackrel{\text{def}}{=} (P_{(i)}, \mathbf{s}_{(i)}, \boldsymbol{\theta}_{(i)})$ (for the ith iteration) by iterative maximization of $\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(i)})$ with

$$\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(i)}) \stackrel{\text{def}}{=} I\!\!E_{\mathfrak{g}_{(i)}}\{\log p_T(Y_{1:T}; X_{1:T}) | Y_{1:T}\}.$$

Namely, one carries out the following two steps:

- (a) E-step: compute $\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(i)})$,
- (b) M-step: choose the update parameters $\mathfrak{g}_{(i+1)} = \arg\max_{\mathfrak{g}} \mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(i)})$.

The essence of the EM algorithm is that $\mathcal{Q}(\mathfrak{g}; \mathfrak{g}_{(i)})$ can be used as a surrogate for $\log p_T(y_1, \ldots, y_T; x_1, \ldots, x_T; \theta)$, see Cappé, Moulines & Rydén (2005).

In our setting, we may write $\mathcal{Q}(\mathfrak{g};\mathfrak{g}_{(i)})$ as:

$$\begin{aligned}\mathcal{Q}(\mathfrak{g};\mathfrak{g}_{(i)}) &= \sum_{i=1}^{M} I\!E_{\mathfrak{g}_{(i)}}[\mathbf{f}\{X_0=i\}\log\{\pi_i f_i(y_0)\}|Y_{1:T}] \qquad (4.11)\\ &\quad+\sum_{t=1}^{T}\sum_{i=1}^{M} I\!E_{\mathfrak{g}_{(i)}}[\mathbf{f}\{X_t=i\}\log f_i(y_t)|Y_{1:T}]\\ &\quad+\sum_{t=1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M} I\!E_{\mathfrak{g}_{(i)}}[\mathbf{f}\{X_t=j\}\mathbf{f}\{X_{t-1}=i\}\log\{p_{ij}\}|Y_{1:T}]\\ &= \sum_{i=1}^{M} I\!P_{\mathfrak{g}_{(i)}}(X_0=i|Y_{1:T})\log\{\pi_i f_i(y_0)\} \qquad (4.12)\\ &\quad+\sum_{t=1}^{T}\sum_{i=1}^{M} I\!P_{\mathfrak{g}_{(i)}}(X_t=i|Y_{1:T})\log f_i(y_t)\\ &\quad+\sum_{t=1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M} I\!P_{\mathfrak{g}_{(i)}}(X_{t-1}=i,X_t=j|Y_{1:T})\log\{p_{ij}\}, \qquad (4.13)\end{aligned}$$

where $f_i(\cdot)$ is as in (4.33) and the margins may be estimated nonparametrically by $\widehat{F}_d^{\mathfrak{m}}(x)=(T+1)^{-1}\sum_{i=1}^{T}\mathbf{f}(X_i\le x)$. The E-step, in which $I\!P_{\mathfrak{g}_{(i)}}(X_t=i|Y_{1:T})$, $I\!P_{\mathfrak{g}_{(i)}}(X_{t-1}=i,X_t=j|Y_{1:T})$ are evaluated, is carried out by the forward-backward algorithm and the M-step is explicit in the p_{ij} and the π_i. Recall that $f_i(\cdot)$ is defined from the last section as $c\{F_1^{\mathfrak{m}}(y_1),F_2^{\mathfrak{m}}(y_2),\ldots,F_d^{\mathfrak{m}}(y_d),s^{(i)},\boldsymbol{\theta}^{(i)}\}f_1^{\mathfrak{m}}(y_1)f_2^{\mathfrak{m}}(y_2)\cdots f_d^{\mathfrak{m}}(y_d)$. Adding constraints to (4.13) yields

$$\mathfrak{L}(\mathfrak{g},\lambda;\mathfrak{g}')=\mathcal{Q}(\mathfrak{g};\mathfrak{g}')+\sum_{i=1}^{M}\lambda_i(1-\sum_{j=1}^{M}p_{ij}) \qquad (4.14)$$

For the M-step, we need to take the first order partial derivative, and plug into (4.14). So, the dependency parameters $\boldsymbol{\theta}$ and the structure parameters $\mathbf{s}$ need to be estimated iteratively, for $\boldsymbol{\theta}^{(i)}$:

$$\frac{\partial\mathfrak{L}(\mathfrak{g},\lambda;\mathfrak{g}')}{\partial\theta_{ij}}=\sum_{t=1}^{T}I\!P(X_t=i|Y_{1:T})\partial\log f_i(y_t)/\partial\theta_{ij}, \qquad (4.15)$$

where, $j=1,\ldots,d-1$. To simplify the procedure, we adopt the HAC estimation method (4.8) with weights in terms of $w_{it}\overset{\text{def}}{=}I\!P(X_t=i|Y_{1:T})$. We also fix $\pi_i, i=$

$1, \ldots, M$ as it influences only the first observation X_0 which may be considered also as given and fixed. The estimation of the transition probabilities p_{ij} follows:

$$\frac{\partial \mathfrak{L}(\mathfrak{g}, \lambda; \mathfrak{g}')}{\partial p_{ij}} = \sum_{t=1}^{T} \frac{I\!P(X_{t-1} = i, X_t = j | Y_{1:T})}{p_{ij}} - \lambda_i \tag{4.16}$$

$$\frac{\partial \mathfrak{L}(\mathfrak{g}, \lambda; \mathfrak{g}')}{\partial \lambda_i} = 1 - \sum_{j=1}^{M} p_{ij}. \tag{4.17}$$

Equating (5.1) and (4.17) yields:

$$\widehat{p}_{i,j} = \frac{\sum_{t=1}^{n} I\!P(X_{t-1} = i, X_t = j | Y_{1:T})}{\sum_{t=1}^{n} \sum_{j=1}^{M} I\!P(X_{t-1} = i, X_t = j | Y_{1:T})} \tag{4.18}$$

4.3 Theoretical Results

Assumptions

A.1 $\{X_t\}$ is stationary and irreducible.

A.2 The family of mixtures of at most M elements $\{f(y, \boldsymbol{\theta}_j, s_j) : \boldsymbol{\theta}_j \in \Theta, s_j \in S\}$ is identifiable w.r.t. the parameters and structures:

$$\sum_{j=1}^{M} \alpha_j f(y, \boldsymbol{\theta}_j, s_j) = \sum_{j=1}^{M} \alpha'_j f(y, \boldsymbol{\theta}'_j, s'_j) \quad a.e. \Longrightarrow \sum_{j=1}^{M} \alpha_j \delta_{s_j} \delta_{\boldsymbol{\theta}_{j,s_j}} = \sum_{j=1}^{M} \alpha'_j \delta_{s'_j} \delta_{\boldsymbol{\theta}'_{j,s'_j}} \tag{4.19}$$

defining δ_{s_j} as the distribution function for a point mass in S, and $\delta_{\boldsymbol{\theta}_{j,s_j}}$ as the distribution function for a point mass in Θ associated with the structure s_j, noting that $\boldsymbol{\theta}_j = \boldsymbol{\theta}'_j$ is only meaningful when $s_j = s'_j$. The property of identifiability is nothing else than the construction of the finite mixture model, McLanchlan & Peel (2000). As a copula is a special form of a multivariate distribution, similar techniques may be applied to get identifiability also in the case of copulae. The family of copula mixtures has been thoroughly investigated in Caia, Chen, Fan & Wang (2006) while developing estimation techniques. In that general case, one should be careful, as the general copula class is very wide and its mixture identification may cause some problems because of the different forms of the densities. The very construction of the HAC narrows this class. Imposing the same generator functions on all levels of the HAC, we restrict the family to the vector of parameters and the tree structure, see also Okhrin et al. (2009). Our preliminary numerical analysis shows that the HAC fulfills the identifiability property for all the structures and parameters used in this study.

A.3 $\{X_t\}_{t=1}^T$ is a time homogeneous Markov chain that is ergodic.

A.4 $\mathbb{E}\{|\log f_i(y, \boldsymbol{\theta}^{(i)}, s^{(i)})|\} < \infty$, for $i = 1, \ldots, M$, $\forall s \in S$.

A.5 For every $\boldsymbol{\theta} \in \Theta$, and any particular structure considered $s \in S$,

$$\mathbb{E}[\sup_{\|\boldsymbol{\theta}' - \boldsymbol{\theta}\| < \delta} \{f_i(Y_1, \boldsymbol{\theta}', s)\}^+] < \infty,$$

for some $\delta > 0$.

Denote as $p_T(y_{1:T}; v, \omega)$ the density in (4.6) with parameters $\{v, \omega\} \in \{V, \Omega\}$ as described in the Appendix 7.2. Define $\widehat{\boldsymbol{\theta}}^{(i)}, \widehat{s}^{(i)}$ as $\widehat{\boldsymbol{\theta}}^{(i)}(\widehat{v}, \widehat{\omega})$ and $\widehat{s}^{(i)}(\widehat{v}, \widehat{\omega})$ with $(\widehat{v}, \widehat{\omega})$ as the point where $p_T(y_{1:T}; v, \omega)$ achieve its maximum value over the parameter space $\{V, \Omega\}$.

It is known that HMM is not itself identifiable as the permutation of states would yield the same value for $p_T(y_{1:T}; v, \omega)$. We assume therefore $\boldsymbol{\theta}^{*(j)}$s to be and $s^{*(j)}$s to be distinct in the sense that for any $s^{*(i)} = s^{*(j)}, i \neq j$ we have $\boldsymbol{\theta}^{*(i)} \neq \boldsymbol{\theta}^{*(j)}$.

Theorem 4.3.1. *Under A.1–A.5, we find the corresponding structure:*

$$\lim_{T \to \infty} \max_{i \in 1, \ldots, M} \mathbb{P}(\widehat{s}^{(i)} = s^{*(i)}) = 1, \forall i. \tag{4.20}$$

Moreover,

Theorem 4.3.2. *Assume A.1–A.5, and $\{Y_t\}_{t=1}^T$ are i.i.d and generated from an HAC HMM model with parameters $\{s^{*(i)}, \theta^{*(i)}, \pi^*, \{p_{ij}^*\}_{i,j}\}$. The parameter $\widehat{\boldsymbol{\theta}}^{(i)}$ satisfies, $\forall \varepsilon > 0$:*

$$\lim_{T \to \infty} \min_{i \in 1, \ldots, M} \mathbb{P}(|\widehat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{*(i)}| > \varepsilon | \widehat{s}^{(i)} = s^{*(i)}) = 0. \tag{4.21}$$

For the proof, we refer to the Appendix.

4.4 Simulation

The estimation performance of HMM HAC is evaluated in this section: subsection I considers four states with very disjoint copulae parameters, while subsection II considers three states realistically calibrated from exchange rate data. We show that our algorithm converges after a few iterations with moderate estimation errors. Throughout the simulation study, we keep the marginal distribution fixed.

4.4.1 Simulation I

In this setup, a three dimensional generating process has fixed marginal distributions: $Y_{t1} \sim \mathrm{N}(0,1)$, $Y_{t2} \sim t(3)$, $Y_{t3} \sim \mathrm{N}(0,3)$. The dependence structure is modeled through HAC with Gumbel generators, and four different dependency parameters and structures corresponding to four states ($M = 4$).

$$
\begin{aligned}
&C\{u_3, C(u_1, u_2; \theta_1 = 4.00); \theta_2 = 1.5\},\\
&C\{u_1, C(u_2, u_3; \theta_1 = 10.0); \theta_2 = 4.0\},\\
&C\{u_2, C(u_1, u_3; \theta_1 = 30.0); \theta_2 = 10.0\},\\
&C\{u_1, C(u_2, u_3; \theta_1 = 40.0); \theta_2 = 20.0\}
\end{aligned}
$$

The quite different state parameters help to easily visualize the dependency states. The transition probability matrix is

$$
P = \{p_{ij}\}_{i,j} = \begin{pmatrix} 0.985 & 0.001 & 0.003 & 0.006 \\ 0.005 & 0.990 & 0.003 & 0.003 \\ 0.005 & 0.005 & 0.991 & 0.001 \\ 0.005 & 0.004 & 0.003 & 0.990 \end{pmatrix}
$$

of sample size $T = 2000$ with $\pi = (0.25, 0.25, 0.25, 0.25)^\top$. Note that we set the diagonal elements of P close to 1, since it is realistic to assume that the states remain the same with a high probability. Figure 4.3 represents the underlying states and a marginal plot of the generated three dimensional time series. No state switching pattern is evident from the marginal plots. Figure 4.4, however, clearly displays the switching of dependency patterns. The black, red, green, and blue dots correspond to the observations from different states. The green points represent the highest correlation state, whereas the red has smaller correlation. The remaining colors blue and black represent states 1 and 2 as described above. One clearly sees that how the HMM changes the dependency structures.

Figure 4.5 displays the first seven iterations. (The parameters remain constant after that). Since the starting values may influence the result, a moving window estimation is proposed to decide the initial parameters. The blue and the red dotted line show, respectively, how the estimators behave with the initial values close to the true (red) and initial values (blue) obtained from the proposed algorithm . The upper panel of Figure 4.5 shows the number of wrongly estimated states at each iteration; the middle panel represents the (L_1) difference of the true transition matrix from the estimated ones; the lower panel is the sum of the estimated parameter errors of the four states with the correctly estimated states. One can see that our choice of initial values can perform as well as the true one.

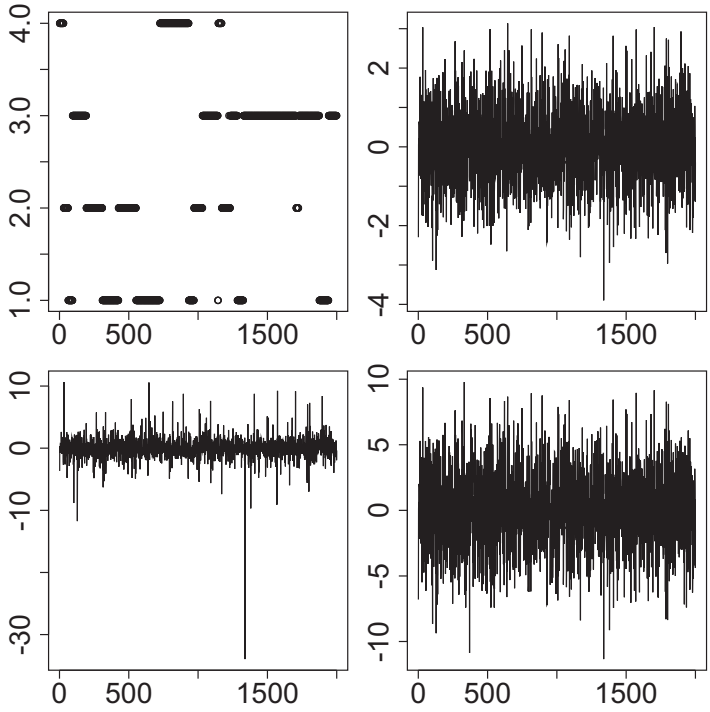


Figure 4.3: The underlying sequence x_t (upper left panel), marginal plots of (y_{t1}, y_{t2}, y_{t3}).

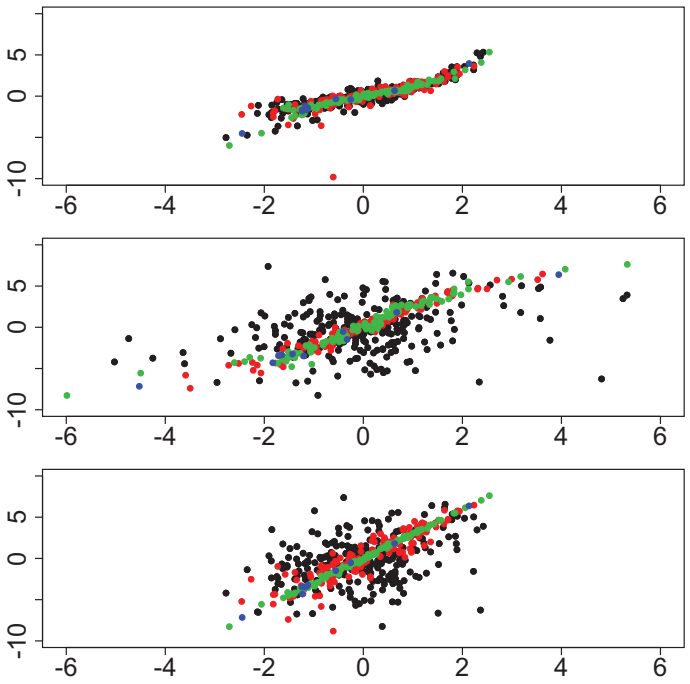


Figure 4.4: Snapshots of pairwise scatter plots of dependency structures ($t = 500, \ldots, 1000$), the (y_{t1}) vs. (y_{t2}) (upper), the (y_{t2}) vs. (y_{t3}) (middle), and the (y_{t1}) vs. (y_{t3})(lower).

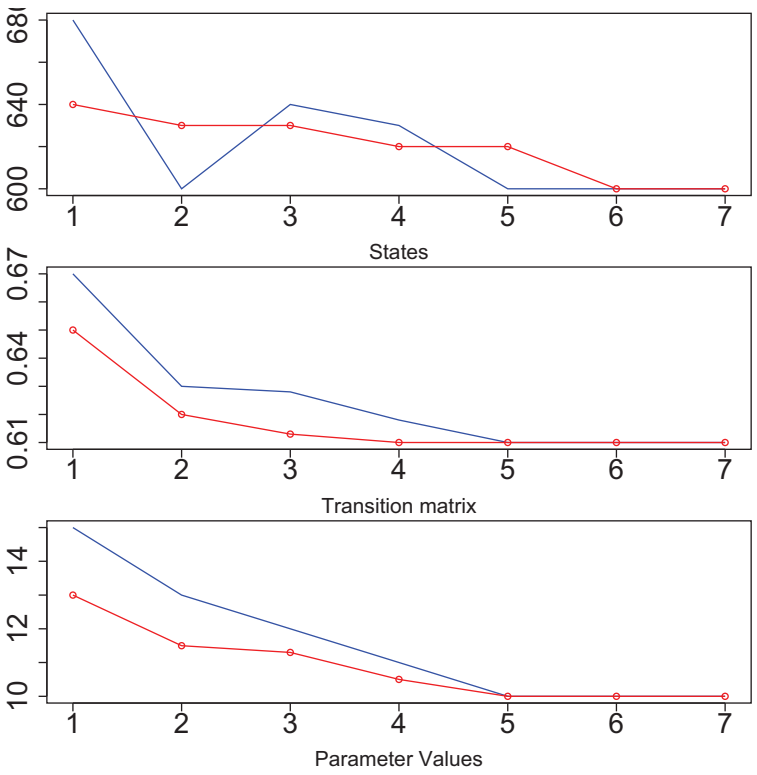


Figure 4.5: The convergence of states (upper panel), transition matrix (middle panel), and parameters (lower panel). Estimation starts from near the true value (red); starts from values provided by our proposal (blue)

4.4.2 Simulation II

Let us consider now a Monte Carlo setup where the setting employs more realistic models. The three states with $M = 3$ are taken as follows:

$$C\{u_1, C(u_2, u_3; \theta_1 = 1.3); \theta_2 = 1.05\}$$
$$C\{u_2, C(u_3, u_1; \theta_1 = 2.0); \theta_2 = 1.35\}$$
$$C\{u_3, C(u_1, u_2; \theta_1 = 4.5); \theta_2 = 2.85\},$$

the transition matrix is chosen as:

$$P = \begin{pmatrix} 0.72 & 0.15 & 0.13 \\ 0.23 & 0.64 & 0.13 \\ 0.03 & 0.02 & 0.95 \end{pmatrix}$$

sample size $T = 2000$. The iteration procedure stops after eleven steps. Figure 4.6 presents the deviations of the estimated states, the transition matrix, and the parameters from their true values. The estimation error is presented in the same fashion as in Figure 4.5. To judge the estimation quality, a histogram of the estimation error from 100 samples is presented in Figure 4.7. The proportion of the misspecified states is centered around roughly $15\% - 17\%$.

4.5 Applications

To see how HMM HAC performs on a real data set, applications to financial and rainfall data are offered. A good model for the dynamics of exchange rates gives insights into exogenous economic conditions, such as the business cycle. It is also helpful for portfolio risk management and decisions on asset allocation. We demonstrate the performance of our proposed technique by applying it to forecasting the VaR of a portfolio and compare it with multivariate GARCH models (DCC, BEKK, etc.) The backtesting results show that the VaR calculated from HMM HAC performs significantly better.
The second application is on modeling a rainfall process. HMM is a conventional model for rainfall data, however, bringing HMM and HAC together for modeling the multivariate a rainfall process is an innovative modeling path.

4.5.1 Application I

Data

The data set consists of the daily values for the exchange rates JPY/EUR, GBP/EUR and USD/EUR. The covered period is [4.1.1999; 14.8.2009], resulting in 2771 ob-

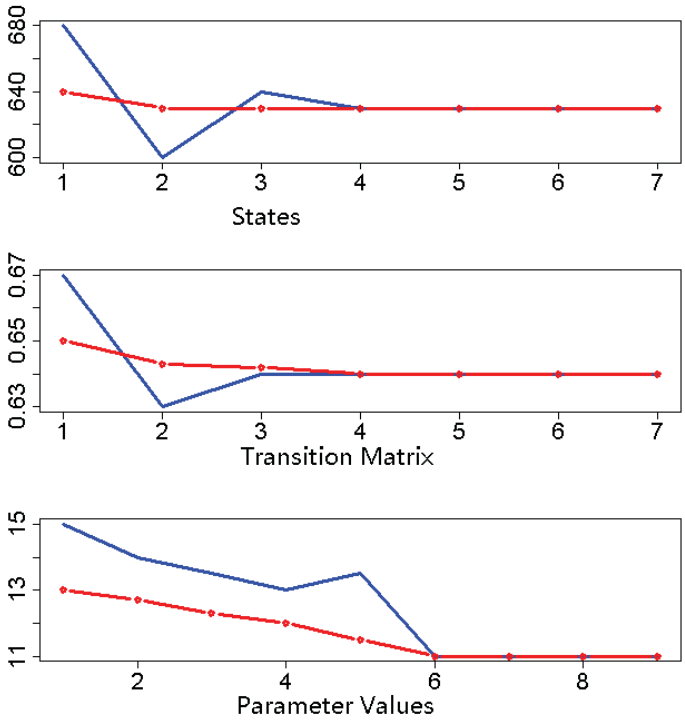


Figure 4.6: The convergence of states (upper panel), transition matrix (middle panel), parameters (lower panel). Estimation starts from near true value (red); starts from values attained by our proposal (blue)

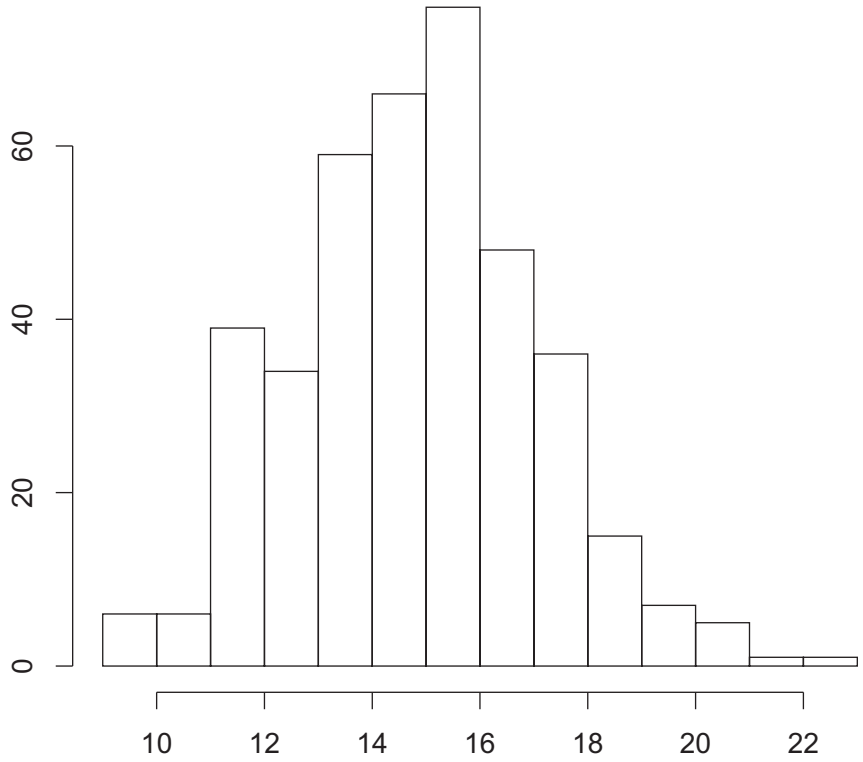


Figure 4.7: The error of misidentification of states from 100 samples

servations.

To eliminate intertemporal conditional heteroscedasticity, we fit to each marginal time series of log-returns a univariate GARCH(1,1) process

$$Y_{j,t} = \mu_{j,t} + \sigma_{j,t}\varepsilon_{j,t} \text{ with } \sigma^2_{j,t} = \omega_j + \alpha_j \sigma^2_{j,t-1} + \beta_j (Y_{j,t-1} - \mu_{j,t-1})^2 \tag{4.22}$$

and $\omega > 0$, $\alpha_j \geq 0$, $\beta_j \geq 0$, $\alpha_j + \beta_j < 1$.

The residuals exhibit the typical behavior: they are not normally distributed, which motivates nonparametric estimation of the margins. From the results of the Box–Ljung test, whose p-values are $0.73, 0.01$, and 0.87 for JPY/EUR, GBP/EUR and USD/EUR, we conclude that the autocorrelation of the residuals is strongly significant only for the GBP/EUR rate. After this intertemporal correction, we work only with the residuals.

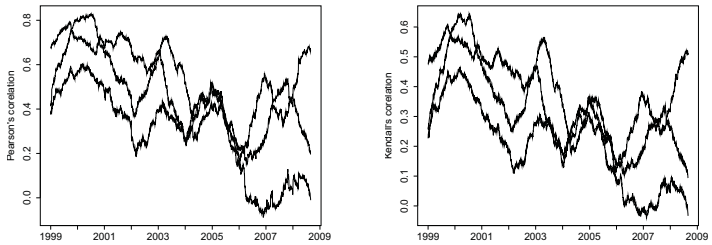


Figure 4.8: Rolling window estimators of Pearson's (left) and Kendall's (right) correlation coefficients between the GARCH(1,1) residuals of exchange rates: JPY and USD (solid line), JPY and GBP (dashed line), GBP and USD (dotted line). The width of the rolling window is set to 250 observations.

The dependency variation is measured by Kendall's and Pearson's correlation coefficients: Figure 4.8 shows the variation of both coefficients calculated in a rolling window of width $r = 250$. Their dynamic behavior is similar, but not identical. This motivates once more a time varying copula based model.

Fitting an HMM model

Figures 4.1, 4.9, and 4.10 summarize the analysis using three methods: moving window, LCP, and HMM HAC. LCP uses moving windows, with varying sizes. To be more specific, LCP is a scaling technique which determines a local homogeneous window at each time point Härdle, Okhrin & Okhrin (2012). In contrast to LCP, HMM HAC is based on a global modeling concept rather than a local one. One observes relatively smooth changes of the parameters, see Figures 4.1 and 4.9. HMM HAC is as flexible as LCP, as can be seen from Figures 4.1, 4.9,

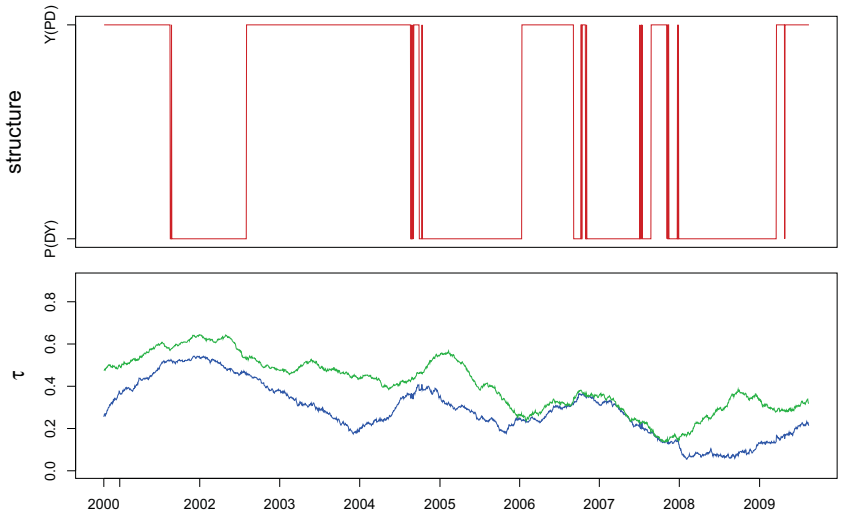


Figure 4.9: Rolling window for exchange rates: structure (upper) and dependency parameters (lower, θ_1 and θ_2) for Gumbel HAC. $w = 250$.

and 4.10, since the structure estimated also takes three values and is confirmed by the variations of structures estimated from LCP. Moreover, the moving window analysis or LCP can serve as a guideline for choosing the initial values for our HMM HAC. Figure 4.11 displays the number of states for HMM HAC for rolling windows with a length of 500 observations.

A VaR estimation example is to show the good performance of HMM HAC. We generate $N = 10^4$ paths with $T = 2219$ observations, and $|W| = 1000$ combinations of different portfolios, where $W = \{(1/3, 1/3, 1/3)^\top \bigcup [\mathbf{w} = (w_1, w_2, w_3)^\top]\}$, with $w_i = w_i' / \sum_{i=1}^3 w_i'$, $w_i' \in U(0, 1)$. The Profit Loss (P&L) function of a weighted portfolio based on assets y_{td} is $L_{t+1} \stackrel{\text{def}}{=} \sum_{d=1}^3 w_i (y_{t+1d} - y_{td})$, with weights $\mathbf{w} = (w_1, w_2, w_3) \in W$. The VaR of a particular portfolio at level $0 < \alpha < 1$ is defined as $VaR(\alpha) \stackrel{\text{def}}{=} F_L^{-1}(\alpha)$, where the $\widehat{\alpha}_{\mathbf{w}}$ is estimated as a relative fraction of violations, see Table 4.1:

$$\widehat{\alpha}_{\mathbf{w}} \stackrel{\text{def}}{=} T^{-1} \sum_{t=1}^{T} \mathbf{I}\{L_t < \widehat{VaR}_t(\alpha)\},$$

and the distance between $\widehat{\alpha}_{\mathbf{w}}$ and α is

$$e_{\mathbf{w}} \stackrel{\text{def}}{=} (\widehat{\alpha}_{\mathbf{w}} - \alpha)/\alpha.$$

If the portfolio distribution is i.i.d., and a well calibrated model is properly mim-

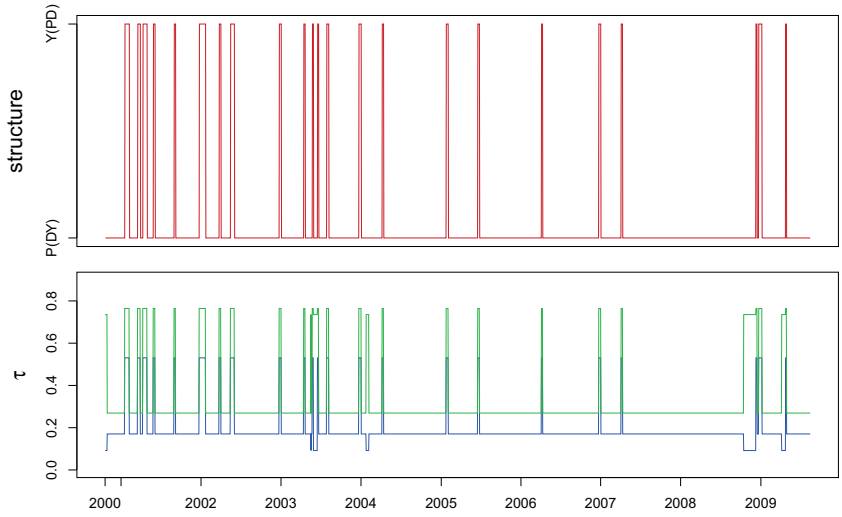


Figure 4.10: HMM for exchange rates: structure (upper) and dependency parameters (lower, θ_1 and θ_2) for Gumbel HAC.

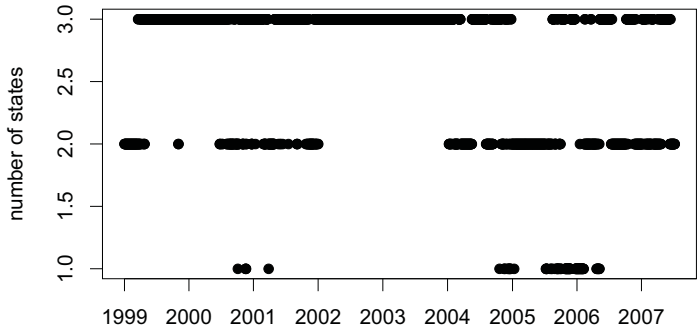


Figure 4.11: Plot of estimated number of states

Window\α		0.1	0.05	0.01
HMM, RGum	500	0.0980	**0.0507**	**0.0128**
HMM, Gum	500	**0.0981**	0.0512	0.0135
Rolwin, RGum	250	0.1037	0.0529	0.0151
Rolwin, Gum	250	0.1043	0.0539	0.0162
LCP, $m_0 = 40$	468	0.0973	0.0520	0.0146
LCP, $m_0 = 20$	235	0.1034	0.0537	0.0169
DCC	500	0.0743	0.0393	0.0163

Table 4.1: VaR backtesting results, $\overline{\widehat{\alpha}}$, where "Gum" denotes the Gumbel copula and "RGum" the rotated Gumbel one.

icking the true underlying asset process, $\widehat{\alpha}_{\mathbf{w}}$ is close to its nominal level α. The performance is measured through an average of $\alpha_{\mathbf{w}}$ over all $|W|$ portfolios, see Table 4.1.

We considered four main models: HMM HAC for 500 observation windows for Gumbel and rotated Gumbel; multiple rolling window with 250 observations windows; LCP with $m_0 = 20$ and $m_0 = 40$ with Gumbel copulae; and DCC, see Engle (2002), based on 500 observation windows. For all the models we made an out of sample forecast. To better evaluate the performance, we calculated the average and SD of e_W:

$$A_W = \frac{1}{|W|} \sum_{\mathbf{w} \in W} e_{\mathbf{w}}, \qquad D_W = \left\{ \frac{1}{|W|} \sum_{\mathbf{w} \in W} (e_{\mathbf{w}} - A_W)^2 \right\}^{1/2}.$$

Tables 4.1 and 4.2 show the backtesting performance for the described models. One concludes that HMM HAC performs better than the concurring moving window, LCP, or DCC, as A_w and D_w are typically smaller.

4.5.2 Application II

A realistic model for rainfall, which can be used to forecast or simulate rainfall is certainly necessary. The difficulty in modeling precipitation data is the nonzero point mass at zero of the rainfall distribution. Another difficulty arises when one incorporates spatial relationships, see Ailliot, Thompson & Thomson (2009) for an HMM application. However, Ailliot et al. (2009) only consider Gaussian dependency among locations, and the method is computationally expensive.

We extend Ailliot et al. (2009) to a copula framework. Different from application I, the marginal distribution here will be varying over states. We propose two

Window\α		0.1	0.05	0.01
HMM, RGum	500	-0.0204 (0.013)	**0.0147** (0.012)	**0.2827** (0.064)
HMM, Gum	500	**-0.0191** (0.008)	0.0233 (0.018)	0.3521 (0.029)
Rolwin, RGum	250	0.0375 (0.009)	0.0576 (0.012)	0.5076 (0.074)
Rolwin, Gum	250	0.0426 (0.009)	0.0772 (0.030)	0.6210 (0.043)
LCP, $m_0 = 40$	468	-0.0270 (0.010)	0.0391 (0.018)	0.4553 (0.037)
LCP, $m_0 = 20$	235	0.0344 (0.009)	0.0735 (0.026)	0.6888 (0.050)
DCC	500	-0.2573 (0.015)	-0.2140 (0.015)	0.6346 (0.091)

Table 4.2: Robustness relative to $A_W(D_W)$

methods for modeling the marginal distributions: one is to take y_{tk} to be censored normal distributions, with the following equation:

$$f_k^{\mathfrak{m}}\{y_{tk}\} = \begin{cases} 1 - p_k^{x_t} & y_{tk} = 0 \\ p_k^{x_t}\varphi[\{y_{tk} - \mu^{x_t}(k)\}/\{\sigma^{x_t}(k)\}]/\sigma^{x_t}(k) & y_{tk} > 0 \end{cases}$$

with $k = 1, \ldots, d$ as the location, $\varphi(\cdot)$ as the standard normal density, $p_k^{x_t}$ as the rainfall occurrence probability for the location k and state x_t, and $\mu^{x_t}(k), \sigma^{x_t}(k)$ the mean and standard deviation parameters at time t for location k.

A second proposal for the marginal distributions are the gamma distributions:

$$f_k^{\mathfrak{m}}\{y_{tk}\} = \begin{cases} 1 - p_k^{x_t} & y_{tk} = 0 \\ p_k^{x_t}\gamma\{y_{tk}; \alpha(k)^{x_t}, \beta(k)^{x_t}\} & y_{tk} > 0, \end{cases}$$

where again the $\alpha(k)^{x_t}, \beta(k)^{x_t}$ are the shape and scale parameters for state x_t and location k. We take the joint distribution function to be a truncated version of a continuous copula function, with the copula density $c_d(\cdot)$ denoted by

$$c_d(\mu, \theta) = \begin{cases} c_c(\mu, \theta), & y_{tk} > 0, \forall k \\ \partial C_c(\mu, \theta)/\partial\mu_{k_1} \ldots \partial\mu_{k_B}, & k_i \in \{y_{tk_i} > 0\}, i \in 1, \ldots, E \end{cases} \tag{4.23}$$

where E denotes the number of wet places among the d locations, the C_c are the continuous copula functions, and c_c are the continuous copula densities. Our formulation is simpler than that of Ailliot et al. (2009) since the copulae have closed-form cdfs, so we do not need additional effort to calculate an integral. The representation in (4.23) is, however, more general, as we consider copulae for capturing the dependencies.

Assume that the daily rainfall observations from the same month are yearly independent realizations of a common underlying hidden Markov model, whose states

Figure 4.12: Map of Guangxi, Guangdong, Fujian in China

represents different weather types. As an example, we take every June's daily rainfall.

$$\begin{aligned}
& \log p_T(y_{1:T}, x_{1:T}; v \times \omega) \\
= \; & \log\{\sum_{i=1}^{M} \mathbf{f}\{x_0 = i\}\pi_i f_i(y_0)\} + \sum_{t=1}^{T} \log\{\sum_{i=1}^{M}\sum_{j=1}^{M} \mathbf{f}\{x_t = j\}\mathbf{f}\{x_{t-1} = i\}p_{ij} f_j(y_t)\} \\
= \; & \sum_{i=1}^{M} \mathbf{f}\{x_0 = i\} \log\{\pi_i f_i(y_0)\} + \sum_{t=1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M} \mathbf{f}\{x_t = j\}\mathbf{f}\{x_{t-1} = i\} \log\{p_{ij} f_j(y_t)\} \\
& + \sum_{t\in B}\sum_{i=1}^{M}\{\mathbf{f}\{x_t = i\}\{\log(\pi_i)\} - \sum_{j=1}^{M} \mathbf{f}\{x_t = j\}\mathbf{f}\{x_{t-1} = i\} \log(p_{ij})\}.
\end{aligned}$$

B is the set of days which are the first day of June for each year. We use here 50 years of rainfall data from three locations in China: Guangxi, Guangdong, and Fujian (Figure 4.12). The graphical correlation can naturally be captured by the fitting of different copulae state parameters.

Table 4.3 presents with a truncated Gumbel the estimated three states, the corresponding different marginal distributions and copula parameters, with estimated initial probability: $\widehat{\pi}_{X_t} = (0.298, 0.660, 0.042)$ and estimated transition probability matrix:

$$\begin{pmatrix} 0.590 & 0.321 & 0.298 \\ 0.188 & 0.742 & 0.660 \\ 0.329 & 0.271 & 0.042 \end{pmatrix}.$$

In our data situation, gamma distributions fit better as marginals. The states

X_t	Shape	Scale	Occur Prob
1	(0.442,0.429,0.552)	(139.33,116.70,169.66)	(0.252,0.256,0.439)
2	(0.671,0.618,0.561)	(273.83,253.25,427.46)	(0.806,0.786,0.683)
3	(0.636,1.125,0.774)	(381.09,264.83,514.08)	(0.667,1.000,0.944)

Table 4.3: Rainfall occurrence probability and shape, scale parameters estimated from HMM (data 1957–2006) .

Location	True	$\widehat{\text{Corr}}(Y_{t,1}, Y_{t,2})$
$1--2$	0.308	0.300 $(0.235, 0.373)$
$2--3$	0.261	0.411 $(0.256, 0.586)$
$1--3$	0.203	0.130 $(0.058, 0.215)$

Table 4.4: True correlations, simulated averaged correlations from 1000 samples their 5% confidence intervals. 1 Fujian, 2 Guangdong, 3 Guangxi

filtered out represents different weather types. The third states are the most humid states, with a high rainfall occurrence probabilities, while the second states are drier, and the first are the driest. From the parameters of the gamma distributions, one sees the variance increases from the first to the third states, which indicates a higher chance for heavy rainfall for the humid states.

To validate our model, 1000 samples of artificial time series of 1500 observations were generated from the fitted model and compared with the original data. Table 4.4 presents the true Pearson correlation compared with the estimated ones from the generated time series. The 5% confidence intervals of the estimators cover the true correlation, which implies that the simulated rainfall can describe the real correlation of the data quite well. Figure 4.13 shows a marginal plot of the log survival function derived from the empirical cdf of the real data and generated data. The log survival function is a transformation of the marginal cdf $F^{\mathfrak{m}}(y_{tk})$:

$$\log\{1 - F^{\mathfrak{m}}(y_{tk})\}. \tag{4.24}$$

Again we show that the 95% confidence interval can cover the true curve fairly well.

Figure 4.14 contains the autocorrelations and cross-correlations of the real data and the generated time series. Unfortunately, our generated time series do not show a similar autocorrelation or cross-correlation. Since there is usually more than one significant lag of autocorrelation or cross-correlation, the simulated time series mostly only have one lag.

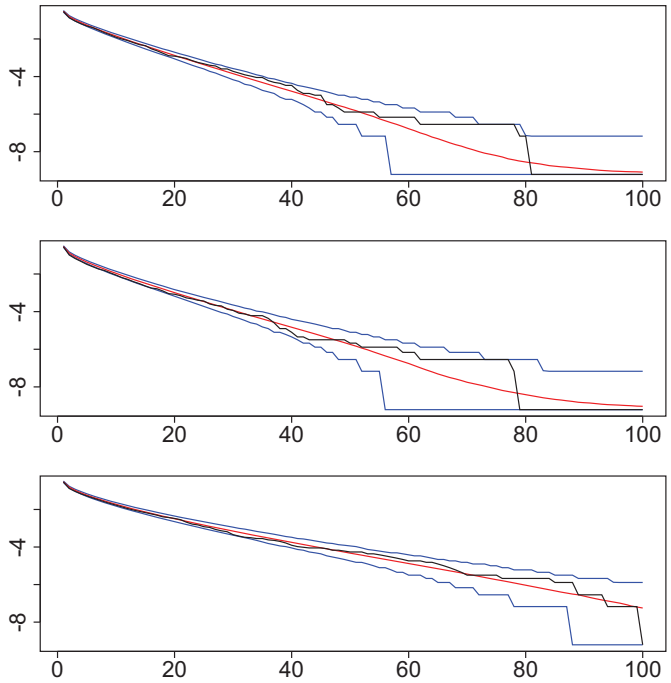


Figure 4.13: Log-survivor-function (red) and 95% prediction intervals (blue) of the simulated distribution for the fitted model with sample log-survivor-function superimposed (black)

4.6 Conclusion

We propose a dynamic model for multivariate time series with non-Gaussian dependency. The idea has an easy extension to HMM for general copula models, and leads to a rich field for further work on dynamic models with dependency structures. This method is helpful in studying financial contagion at an extreme level over time, and naturally it can help in deriving conditional risk measures, such as CoVaR. As we have shown, dynamic copula models are good enough to mimic financial markets as well as nature.

4.7 Appendix

4.7.1 Copulae

Let $Z_1, \ldots, Z_d$ be r.v. with continuous cumulative distribution function (cdf) $F(\cdot)$. The Sklar theorem guarantees the existence and uniqueness of copula functions by stating that there exists a unique function $C : [0,1]^d \to [0,1]$ satisfying

$$C(u_1, \ldots, u_d) = F\{F_1^{-1,\mathfrak{m}}(u_1), \ldots, F_d^{-1,\mathfrak{m}}(u_d)\}, \quad u_1, \ldots, u_d \in [0,1],$$

where $F_1^{-1,\mathfrak{m}}(u_1), \ldots, F_d^{-1,\mathfrak{m}}(u_d)$ are the quantile functions of the corresponding continuous marginal distributions $F_1^{\mathfrak{m}}(Z_1), \ldots, F_d^{\mathfrak{m}}(Z_d)$.

One of the families, which are flexible enough to capture a tail dependency, have an explicit form, and are simple to estimate is the family of Archimedean copulae, see Nelsen (2006).

$$C(u_1, \ldots, u_k) = \phi\{\phi^{-1}(u_1) + \cdots + \phi^{-1}(u_d)\}, \quad u_1, \ldots, u_d \in [0,1], \tag{4.25}$$

where $\phi(.)$ is defined as the generator of the copula and depends on the parameter θ. $\phi(.) \in \mathfrak{L} = \{\phi(.) : [0;\infty) \to [0,1] \,|\, \phi(0) = 1,\ \phi(\infty) = 0;\ (-1)^j \phi^{(j)} \geq 0;\ j = 1, \ldots, \infty\}$; simplified assumptions on ϕ may be found in McNeil & Nešlehová (2009). As an example, the Gumbel generator is given by $\phi(.) = \exp(-x^{1/\theta})$ for $0 \leq x < \infty,\ 1 \leq \theta < \infty$.

In this work we consider less restrictive compositions of simple Archimedean copulae leading to a Hierarchical Archimedean Copula (HAC) $C(u_1, \ldots, u_d; s, \boldsymbol{\theta})$, where $s = \{(\ldots(i_1 \ldots i_{j_1}) \ldots (\ldots) \ldots)\}$ denotes the structure of HAC, with $i_\ell \in \{1, \ldots, d\}$ being a reordering of the indices of the variables and s_j the structure of the sub-copulae with $s_d = s.$ and $\boldsymbol{\theta}$ is the set of copula parameters. For example, the fully nested HAC (see Figure 4.15, left) can be expressed by

$$\begin{aligned} & C(u_1, \ldots, u_d; s = s_d, \boldsymbol{\theta}) = C\{u_1, \ldots, u_d; ((s_{d-1})d), (\theta_1, \ldots, \theta_{d-1})^\top\} \\ = \ & \phi_{d-1,\theta_{d-1}}(\phi^{-1}_{d-1,\theta_{d-1}} \circ C\{u_1, \ldots, u_{d-1}; ((s_{d-2})(d-1)), (\theta_1, \ldots, \theta_{d-2})^\top\} + \phi^{-1}_{d-1,\theta_{d-1}}(u_d)), \end{aligned}$$

where $s = \{(\dots(12)3)\dots)d)\}$. On the RHS of Figure 4.15 we have the partially nested HAC with $s = ((12)(34))$ in dimension $d = 4$. For more details of HAC, see Joe (1997), Whelan (2004), Savu & Trede (2006), Okhrin et al. (2009).

Not all generator functions can be mixed within one HAC. To make the problem more concrete, we concentrate on one single generator family within one HAC, and the discussion is constrained to binary structures, i.e., at each level of the hierarchy only two variables are joined together. This makes our model very flexible and yet also parsimonious.

Note that for each HAC not only are the parameters unknown, but also the structure has to be determined. We adopt the computation procedure in Okhrin et al. (2009) to estimate the HAC structure and parameters, which leads to efficient and unbiased estimators. In this procedure, one estimates the marginal distributions either parametrically or nonparametrically. Then assuming that the marginal distributions are known, one selects the couple of variables with the strongest fit and denotes the corresponding estimator of the parameter at the first level by $\widehat{\theta}_1$ and the set of indices of the variables by I_1. The selected couple is joined together to define the pseudo-variables $z_1 = C\{(I_1); \widehat{\theta}_1, \phi_1\}$. Next, one proceeds in the same way by considering the remaining variables and the new pseudo-variable. At every level, the copula parameter is estimated by assuming that the margins as well as the copula parameters at lower levels are known. This procedure allows us to determine the estimated structure of the copula recursively.

4.7.2 Proof of Theorems 4.3.1 and 4.3.2

In the HMM HAC framework, let $\{X_t, t \geq 0\}$ with transition probability matrix $P^{v,\omega} = [p_{ij}^{v,\omega}]_{i,j=1,\dots,M}$ and initial distribution $\pi^{v,\omega} = \{\pi_i^{v,\omega}\}_{i=1,\dots,M}$, where $\{v, \omega\} \in \{V, \Omega\} \subseteq \{N^*, \mathbb{R}^q\}$ denotes an element in the parameter space $\{V, \Omega\}$ which parametrizes this model, and q is the number of continuous parameters (note that our parameter space is partly discrete (V), and partly continuous (Ω)). Suppose that a real-valued additive component $B_{t,j} = \sum_{k=0}^{t} Y_{k,j}, j \in 1, \dots, d$, with $B_t = (B_{t,1}, B_{t,2}, \dots, B_{t,d})^\top$ and with $Y_k = (Y_{k,1}, Y_{k,2}, \dots, Y_{k,d})^\top$ a r.v. taking values on $\mathbb{R}^d$, is adjoined to the chain such that $\{(X_t, B_t), t \geq 0\}$ is a Markov chain on $D \times \mathbb{R}^d$ and

$$\begin{aligned} & I\!P\{(X_t, B_t) \in A \times (B + b) | (X_{t-1}, B_{t-1}) = (i, b)\} \qquad (4.26) \\ = \; & I\!P\{(X_1, B_1) \in A \times B | (X_0, B_0) = (i, 0)\} \\ = \; & I\!P(i, A \times B) = \sum_{j \in A} \int_{b \in B} p_{ij}^{v \times \omega} f_j\{b; s^{(j)}(v \times \omega), \boldsymbol{\theta}^{(j)}(v \times \omega)\} \mu(db), \end{aligned}$$

where $B, b \subseteq \mathbb{R}^d$, $A \subseteq D$, $f_j\{b; s^{(j)}(v,\omega), \boldsymbol{\theta}^{(j)}(v,\omega)\}$ is the conditional density of Y_t given X_{t-1}, X_t with respect to a σ-finite measure μ on $\mathbb{R}^d$, and $\boldsymbol{\theta}(v,\omega) \in \Theta, s(v,\omega) \in S$, $j = 1, \ldots, M$ are the unknown parameters. That is, $\{X_t, t \geq 0\}$ is a Markov chain, given $X_0, X_1, \ldots, X_T$, with $Y_1, \ldots, Y_T$ being independent. We give a formal definition. $\{B_t, t \geq 0\}$ is called a hidden Markov model if there is a Markov chain $\{X_t, t \geq 0\}$ such that the process $\{(X_t, B_t), t \geq 0\}$ satisfies (4.26). Note that in (4.26), the usual parameterization $\boldsymbol{\theta}^{(j)}(v,\omega) = \boldsymbol{\theta}^{(j)}$, and $s^{(j)}(v,\omega) = s^{(j)}$. Moreover, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)})^\top \in \mathbb{R}^{dM}$ are the unknown dependency parameters, $\mathbf{s} = (s^{(1)}, \ldots, s^{(M)})$ are the unknown structure parameters, and its true value is denoted by $\boldsymbol{\theta}^*$ and $\mathbf{s}^*$. For simplicity, we will use π_i for $\pi_i^{v,\omega}$ and p_{ij} for $p_{ij}^{v,\omega}$. See Figure 4.2 for a graphical illustration.

Recall the associated parameter space $\{V, \Omega\}$, where V consists of a set of discrete finite elements and Ω is associated with the parameters $\boldsymbol{\theta}, [p_{ij}]_{i,j}$. Define $\mathbf{s}^*$ and $\boldsymbol{\theta}^*$ associated with the point $\{v^0, \omega^0\}$ in the parameter space, as in the following definitions:

$$q_T(Y_{1:T}; v^0, \omega^0) \stackrel{\text{def}}{=} \max_{j \in 1,\ldots,M} p_T(Y_{1:T} | x_1 = j, ; v^0, \omega^0) \tag{4.27}$$

$$H(v^0, \omega^0) \stackrel{\text{def}}{=} I\!E_{v^0,\omega^0}\{-\log p(Y_0 | Y_{-1}, Y_{-2}, \ldots; v_0, \omega_0)\},$$

where $Y_{-1}, \ldots, Y_{-T}$ are a finite number of past values of the process.

$$H(v^0, \omega^0, v, \omega) \stackrel{\text{def}}{=} I\!E_{v^0,\omega^0}\{\log p_T(Y_{1:T}; v, \omega)\}$$

Theorem 4.7.1 (Leroux (1992))**.** *Under A.1–A.5,*

$$\begin{aligned} \lim_{T\to\infty} T^{-1} I\!E_{v^0,\omega^0}\{\log p_T(Y_{1:T}; v^0, \omega^0)\} &= -H(v^0, \omega^0) \\ \lim_{T\to\infty} T^{-1} \log p_T(Y_{1:T}; v^0, \omega^0) &= -H(v^0, \omega^0), \end{aligned}$$

with probability 1*, under* (v^0, ω^0)*, and*

$$\begin{aligned} \lim_{T\to\infty} T^{-1} I\!E_{v^0,\omega^0}\{\log p_T(Y_{1:T}; v, \omega)\} &= H(v^0, \omega^0, v, \omega) \\ \lim_{T\to\infty} T^{-1} \log p_T(Y_{1:T}; v, \omega) &= H(v^0, \omega^0, v, \omega), \end{aligned}$$

with probability 1*, under* (v_0, ω_0)*.*

Lemma 3. *$\forall v_i, u_j, i, j \in 1, \ldots, M$ as weights, the difference between M linear combination of states would lead to*

$$\sum_{i=1}^{M} v_i f(y, s^{(i)}, \boldsymbol{\theta}_{s_{(i)}}) \neq \sum_{j=1}^{M} \mu_j f(y, s^{(j)}, \boldsymbol{\theta}_{s_{(j)}}). \tag{4.28}$$

Proof. For each $s^{(i)}, i \in 1, \ldots, M$ associated with dependency parameter $\boldsymbol{\theta}_{s^{(i)}} \in \mathbb{R}_+^d$.
So

$$\sum_{i=1}^{M} v_i \delta_{s^{(i)}} \neq \sum_{j=1}^{M} \mu_j \delta_{s^{(j)}}, a.e. \tag{4.29}$$

implies

$$\sum_{i=1}^{M} v_i \delta_{s^{(i)}} \delta_{\boldsymbol{\theta}_{s^{(i)}}} \neq \sum_{j=1}^{M} \mu_j \delta_{s^{(j)}} \delta_{\boldsymbol{\theta}_{s^{(j)}}}, a.e..$$

□

Also if (4.29), then the corresponding point in the parameter space (v, ω) would lead to $\mathcal{K}(v, \omega; v^0, \omega^0)$, and (v, ω) would not be in the equivalent class of (v^0, ω^0) as long as the point v and v^0 are different as (4.29) , (the equivalence class of v^0 is defined in Leroux (1992)), and

$$\mathcal{K}(v, \omega; v^0, \omega^0) \stackrel{\text{def}}{=} \int \sum_j u_j p_2(y_1, y_2|j, v_0, \omega_0) \log\{\frac{\sum_j u_j p_2(y_1, y_2|j, v_0, \omega_0)}{\sum_j v_j p_2(y_1, y_2|j, v, \omega)}\} d\mu(y_1) d\mu(y_2) dQ(\mu, v),$$

with $Q(\mu, v)$ as the distribution of $P_(X_1 = j|Y_{-1}, \ldots, Y_{-\infty}), j \in 1, \ldots, M$ under the true measure corresponding to (v_0, ω_0) . Then it follows from A.2 that (4.29) implies $\mathcal{K}(v, \omega; v^0, \omega^0) > 0$.

Next, we study whether plugging in nonparametric estimated margins would affect the consistency results by analyzing the uniform convergence of $\widehat{f}(y, \boldsymbol{\theta}_j, s_j)$.

As $\widehat{f}(y, \boldsymbol{\theta}_j, s_j) = \widehat{c}\{F_1^{\mathfrak{m}}(y_1), F_2^{\mathfrak{m}}(y_2), \ldots, F_d^{\mathfrak{m}}(y_d), \boldsymbol{\theta}^{(i)}, s^{(i)}\} \widehat{f}_1^{\mathfrak{m}}(y_1) \widehat{f}_2^{\mathfrak{m}}(y_2) \cdots \widehat{f}_d^{\mathfrak{m}}(y_d)$

We have the uniform consistency of copulae density,

$$\sup_{u_1, \ldots, u_d \in [0,1]^d} |\widehat{c}(u_1, u_2, \ldots, u_d) - c(u_1, u_2, \ldots, u_d)| = \mathcal{O}_p(T^{-1/2} \log T^{1/2}) \tag{4.30}$$

and according to Bickel & Rosenblatt (1973),

$$\sup_{x \in B} |\widehat{f}_i^{\mathfrak{m}}(x) - f_i^{\mathfrak{m}}(x)| = \mathcal{O}((Th)^{-1/2} \log T^{1/2}) \tag{4.31}$$

Therefore,

$$\sup_{y \in B^d} |\widehat{f}(y, \boldsymbol{\theta}_j, s_j) - f(y, \boldsymbol{\theta}_j, s_j)| = \mathcal{O}((Th)^{-1/2} \log T^{1/2}).$$

So the plug in estimation would not contaminate the consistency results.

To prove the consistency of our estimation of this parameter, we restate the theorems of consistency in Leroux (1992) for our parameter space. One needs to show that first for the discrete subspace V^c which does not contain any point of the equivalence class of v^0 , for $v \in V^c$ and any arbitrary value of $\omega \in \Omega$, it holds, with probability 1,

$$\lim_{T\to\infty} \max_{v\in V^c} \log \sup_{\omega\in\Omega} p_T(Y_{1:T}; v, \omega) - \log p_T(Y_{1:T}; v^0, \omega^0) \to -\infty. \tag{4.32}$$

The fact follows directly from lemma 3 (the identifiability of the states parameters), and its consequence $\mathcal{K}(v, \omega; v^0, \omega^0) > 0$. Theorem 4.3.1 is proved.

To prove Theorem 4.3.2, note that $\lim_{T\to\infty} \min_{i\in 1,\ldots,M} I\!P(|\widehat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{*(i)}| > \varepsilon | \widehat{s}^{(i)} = s^{*(i)})$ is conditioning on the event $\{\widehat{s}^{(i)} = s^{*(i)}\}$ which asymptotically holds with probability 1. Therefore it is suffice to prove, for any $\widehat{s}^{(i)} = s^{(i)}$

$$\lim_{T\to\infty} \min_{i\in 1,\ldots,M} I\!P(|\widehat{\boldsymbol{\theta}}^{(i)} - \boldsymbol{\theta}^{*(i)}| > \varepsilon) = 0. \tag{4.33}$$

To show (4.33), one needs to show that for (V^c, Ω^c) which does not contain any point of the equivalence class of (v^0, ω^0), we have, with probability 1,

$$\lim_{T\to\infty} \{\log \sup_{\omega\in\Omega^c} p_T(Y_{1:T}; v^0, \omega) - \log p_T(Y_{1:T}; v^0, \omega^0)\} \to -\infty, \tag{4.34}$$

which is implied from the following statement: for any closed subset C of Ω^c, there exists a sequence of open subsets of $\mathcal{O}_{\omega_h}$ with $h = 1, \ldots, H$ with $C \subseteq \cup_{h=1}^{H} \mathcal{O}_{\omega_h}$, such that

$$\lim_{T\to\infty} \{\max_h \log \sup_{\omega\in\mathcal{O}_{\omega_h}} p_T(Y_{1:T}; v^0, \omega) - \log p_T(Y_{1:T}; v^0, \omega^0)\} \to -\infty. \tag{4.35}$$

To prove (4.35), we have the modified definition:

$$H(v^0, \omega^0, v^0, \omega; \mathcal{O}_{\omega_h}) \stackrel{\text{def}}{=} \lim_T \log \sup_{\omega'\in\omega^0} q_T(Y_{1:T}, v^0, \omega')/T. \tag{4.36}$$

It can be derived that

$$H(v^0, \omega^0, v^0, \omega) < H(v^0, \omega^0, v^0, \omega^0), \tag{4.37}$$

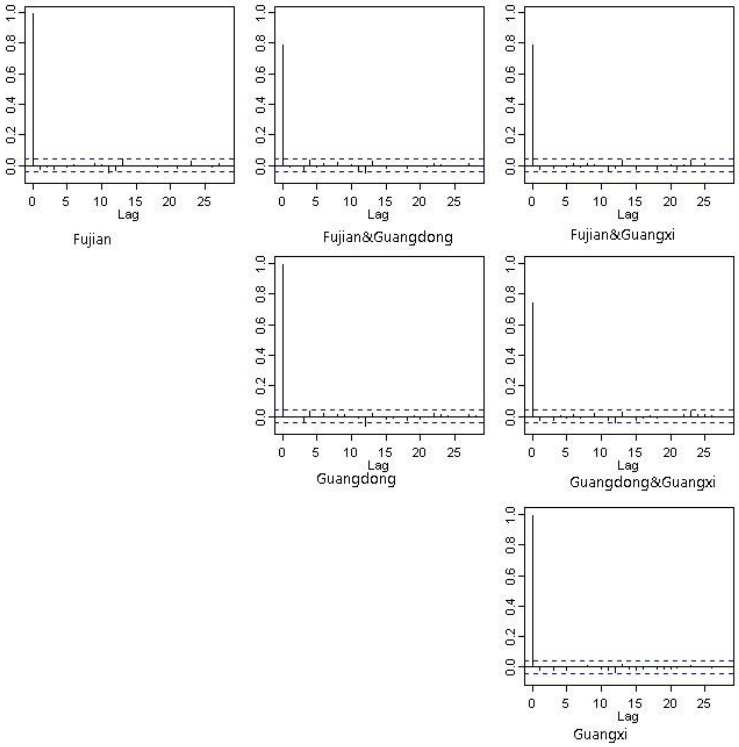
for (v^0, ω) and (v^0, ω^0) does not lie in the same equivalence class. Then (4.37) is a consequence of the identifiability condition A.2, and this leads to: $\exists \varepsilon > 0, T_\varepsilon$ and $\mathcal{O}_\omega$ such that

$$I\!E \log \sup_{\omega'\in\mathcal{O}_\omega} q_{T_\varepsilon}(v^0, \omega')/T_\varepsilon < I\!E \log q_{T_\varepsilon}(v^0, \omega)/T_\varepsilon + \varepsilon < H(v^0, \omega^0, v^0, \omega^0) - \varepsilon.$$

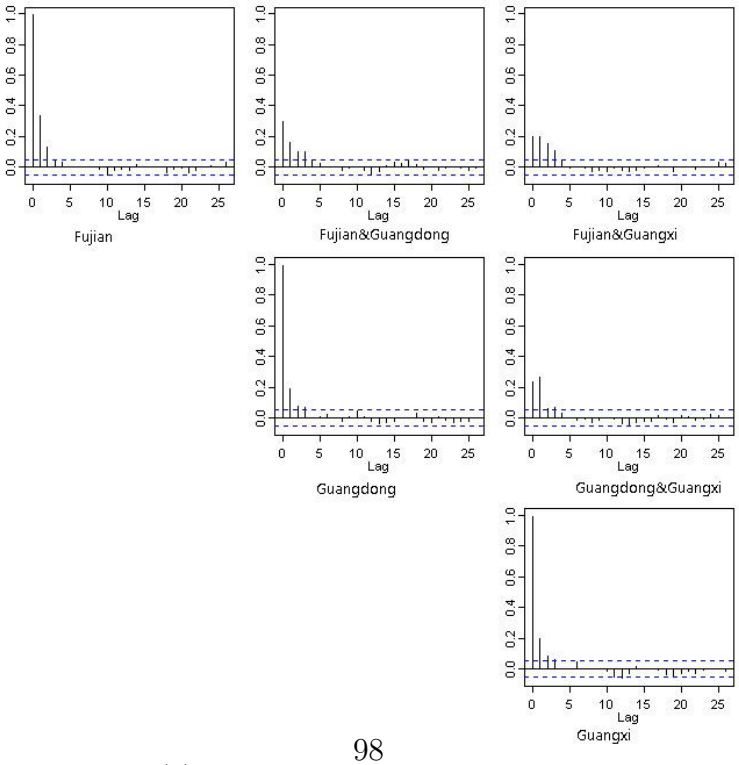
Also because $\log \sup_{\omega' \in \mathcal{O}_\omega} p_T(Y_{1:T}, v^0, \omega')/T$ and $\log \sup_{\omega' \in \mathcal{O}_\omega} q_T(Y_{1:T}, v^0, \omega')/T$ have the same limit value, there exists a constant $\varepsilon > 0$,

$$\lim_{T \to \infty} \log \sup_{\omega' \in \mathcal{O}_{\omega_h}} p_T(y_{1:T}, v^0, \omega')/T = H(v^0, \omega^0, v^0, \omega; \mathcal{O}_{\omega_h}) \leq H(v^0, \omega^0, v^0, \omega^0) - \varepsilon.$$

Now (4.35) follows.



(a) the simulated rainfall time series.



(b) the original rainfall time series.

Figure 4.14: Autocorrelations and cross-correlations of the simulated rainfall and original time series

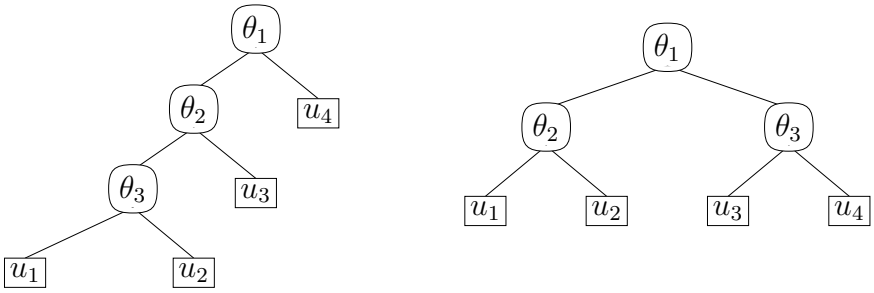


Figure 4.15: Fully and partially nested copulae of dimension $d = 4$ with structures $s = (((12)3)4)$ on the left and $s = ((12)(34))$ on the right

Chapter 5

Localising temperature risk

5.1 Introduction

Pricing of contingent claims based on stochastic dynamics for example stocks or FX rates is well known in financial engineering. An elegant access of such a pricing task is based on self-financing replication arguments. An essential element of this approach is the tradability of the underlying. This however does not apply to weather derivatives contingent on temperature or rain since the underlying is not tradable. In this context, the proposed pricing techniques are based on either equilibrium ideas (Horst & Mueller (2007)) or econometric modelling of the underlying dynamics Campbell & Diebold (2005) and Benth, Benth & Koekebakker (2007) followed by risk neutral pricing.

The equilibrium approach relies on assumptions about preferences (with explicitly known functional forms) though. In this study we prefer a phenomenological approach since the underlying (temperature) we consider is of local nature and our analysis aims at understanding the pricing at different locations and different time points around the world. Such a time series approach has been taken by Benth et al. (2007), who corrects for seasonality (in mean), then for intertemporal correlation and finally as in Campbell & Diebold (2005), for seasonal variation in volatility. After these manipulations, a Gaussian risk factor needs to be isolated in order to apply continuous time pricing techniques, Karatzas & Shreve (2001).

Empirical studies following this econometrical route show evidence that the resulting risk factor deviates severely from Gaussianity, which in turn challenges the pricing tools, Benth, Härdle & López Cabrera (2011). In particular, for Asian cities, like for example Kaohsiung (Taiwan), one observes very distinctive non-normality in the form of clearly visible heavy tails caused by extended volatility in peak seasons. This is visible from Figure 5.1 where a log density plot reveals a

nonnormal shoulder structure (kurtosis= 3.22, skewness= -0.08, JB= 128.74).

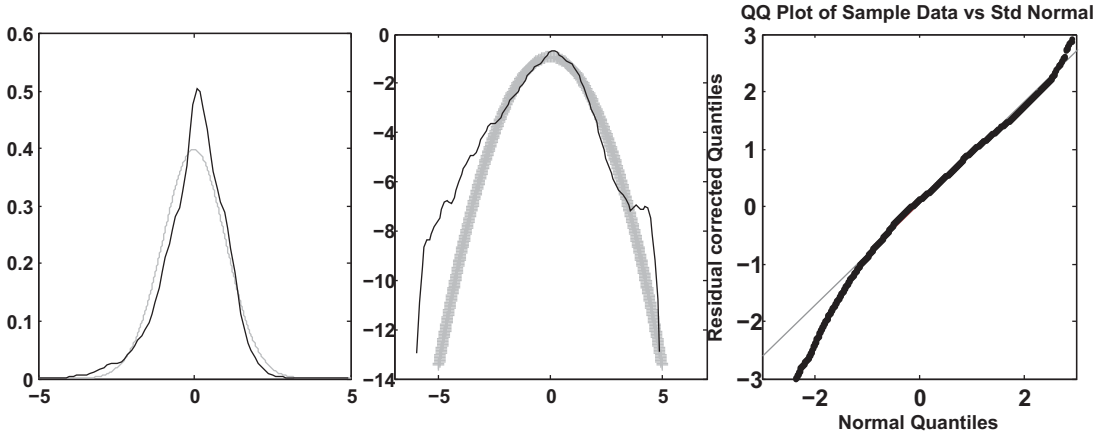


Figure 5.1: Kernel density estimates (left panel), Log normal densities (middle panel) and QQ-plots (right panel) of normal densities (gray lines) and Kaohsiung standardised residuals (black line)

As in Benth et al. (2007) temperature T_t is decomposed into a seasonality term Λ_t and a stochastic part with seasonal volatility σ_t.

The fitted seasonality trend Λ_t and seasonal variance σ_t^2 are approximated with Fourier series (and an additional GARCH term):

$$\Lambda_t = a + bt + \sum_{l=1}^{L} c_l \cos\left\{\frac{2\pi(t-d_l)}{l\cdot 365}\right\}, \tag{5.1}$$

$$\sigma^2_{t,FTSG} = c_{10} + \sum_{l=1}^{L}\left\{c_{2l}\cos\left(\frac{2l\pi t}{365}\right) + c_{2l+1}\sin\left(\frac{2l\pi t}{365}\right)\right\} \tag{5.2}$$

$$+\alpha_1(\sigma_{t-1}\eta_{t-1})^2 + \beta_1\sigma^2_{t-1}, \tag{5.3}$$

$$\eta_t \sim iid(0,1).$$

The upper panel of Figure 5.2 displays the seasonality and deseasonalised residuals over two years in Kaohsiung. The lower panel RHS displays the empirical and seasonal variance function, while the lower panel LHS shows the smoothed seasonal variance function over years. The series expansion (5.1), (5.3) failed though in the volatility peak seasons. Even incorporating an asymmetry term for the dip of temperature in winter does not improve the closeness to normality.

One may of course pursue a fine tuning of (5.1) and (5.3) with more and more periodic terms but this will increase the number of parameters. We therefore propose a local parametric approach. The seasonality Λ_s and σ_s are approximated with a Local Linear Regression (LLR) estimator:

$$\arg\min_{e,f} \sum_{t=1}^{365} \left\{ \overline{T}_t - e_s - f_s(t-s) \right\}^2 K\left(\frac{t-s}{h} \right) \tag{5.4}$$

$$\arg\min_{g,v} \sum_{t=1}^{365} \left\{ \widehat{\varepsilon}_t^2 - g_s - v_s(t-s) \right\}^2 K\left(\frac{t-s}{h} \right) \tag{5.5}$$

where $\overline{T}_t$ is the mean (over years) of daily averages temperatures, $\widehat{\varepsilon}_t^2$ the squared residual process (after seasonal and intertemporal fitting), h the bandwidth and $K(\cdot)$ is a kernel. Note, that due to the spherical character of the data, the kernel weights in (5.4), (5.5) may be calculated from "wrapped around observations" thereby avoiding bias. The estimates $\widehat{\Lambda}_s$, $\widehat{\sigma}_s^2$ are given by the minimizers $\widehat{e}_s$, $\widehat{g}_s$ of (5.4), (5.5). The upper panel of Figure 5.2 shows the seasonality in mean and the bottom panel on the RHS the volatility estimated with Fourier series and local linear regression using the quartic kernel. We observe high variance in winter and early summer and low variance in spring and late summer.

The scale correction of the obtained residuals (after seasonal and intertemporal fitting) is apparently not identical over the year. A very structured volatility pattern up to April is followed by a moderately constant period until an increasing peak starting in September. This motivates our research to localise temperature risk. The local smoothness of σ_t^2 is of course not only a matter of one location (here Kaohsiung) but varies also over the different cities around the world that we are analysing in this study. Our study is local in a double sense: local in time and space. We use adaptive methods to localise the underlying dynamics and with that being able to achieve Gaussian risk factors. This will justify the pricing via standard tools that are based on Gaussian risk drivers. The localisation in time is based on adjusting the smoothing parameter h. For a general framework on local parametric approximation we refer to Spokoiny (2009). As a result we obtain better approximations to normality and therefore less biased prices.

This chapter is structured as follows. Section 2 describes the localising approach.

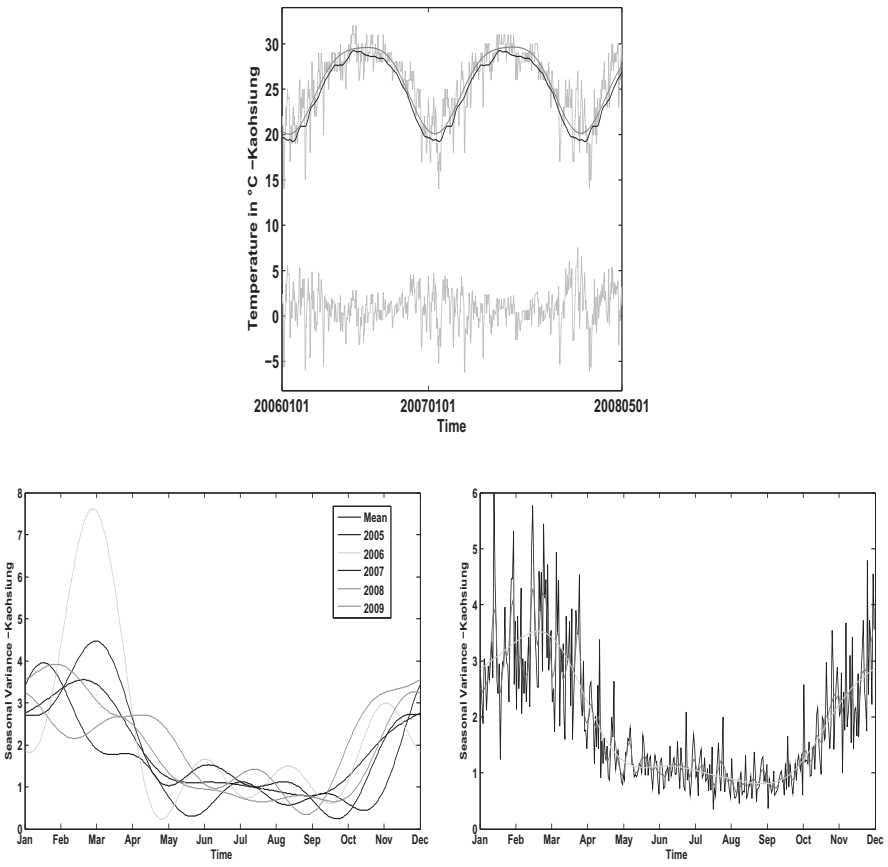


Figure 5.2: Upper panel: Kaohsiung daily average temperature (black line), Fourier truncated (dotted gray line) and local linear seasonality function (gray line), Residuals in lower part. Lower left panel: Fourier seasonal variation ($\widehat{\Lambda}_t$) over time. Lower right panel: Kaohsiung empirical (black line), Fourier (dotted gray line) and local linear (gray line) seasonal variance ($\widehat{\varepsilon}_t^2$) function.

In section 3, we present the data and conduct the analysis to different cities. Section 4 presents an application where the pricing of weather derivative contract types is presented. Section 5 concludes the chapter. All quotations of currency in this chapter will be in USD and therefore we will omit the explicit notion of the currency. All the CAT bond computations were carried out in Matlab version 7.6 and R. The temperature data for different cities in US, Europe and Asia were obtained from the National Climatic Data Center (NCDC), the Deutscher Wetterdienst (DWD), Bloomberg Professional Service and the Japanese Meteorological Agency (JMA).

5.2 Model

Let us change our notation from $t \mapsto (t, j)$, with $t = 1, \ldots, \tau = 365$ days, $j = 0, \ldots, J$ years. The time series decomposition we consider is given as:

$$\begin{aligned} X_{365j+t} &= T_{t,j} - \Lambda_t, \\ X_{365j+t} &= \sum_{l=1}^{L} \beta_{lj} X_{365j+t-l} + \varepsilon_{t,j}, \\ \varepsilon_{t,j} &= \sigma_t e_{t,j}, \\ e_{t,j} &\sim \mathrm{N}(0,1), \\ \widehat{\varepsilon}_{t,j} &= X_{365j+t} - \sum_{l=1}^{L} \widehat{\beta}_{lj} X_{365j+t-l}, \end{aligned} \tag{5.6}$$

where $T_{t,j}$ is the temperature at day t in year j, Λ_t denotes the seasonality effect and σ_t the seasonal volatility. Motivation of this modeling approach can be found in Diebold & Inoue (2001). Later studies like e.g. Campbell & Diebold (2005) and Benth et al. (2007) have provided evidence that the parameters β_{lj} are likely to be j independent and hence estimated consistently from a global autoregressive process model $AR(L_j)$ with $L_j = L$. Since the stylised facts of temperature are reoccurring every year, our focus is on flexible estimation of Λ_t and σ_t^2, see Figure 5.2.

The seasonal trend function Λ_t and the seasonal variance function σ_t^2 affect the Gaussianity of the resulting normalised residuals. The commonly used approaches 1. truncated Fourier series, 2. local polynomial regression are both too restrictive and do not fit the data well since they are not yielding normal risk factors. These observations motivate us to consider a more flexible approach. The main idea is to fit a simple parametric model locally for the trend and variance with adaptively chosen window sizes. Specifically, we use kernel smoothing and adopt an adaptive technique to choose the bandwidth over days. Other examples of this technique

can be found in Cízek, Härdle & Spokoiny (2009) and Chen, Härdle & Pigorsch (2010).

5.2.1 How does the adaptation work?

The time series $T_{t,j}$ are approximated at a fixed time point $s \in [1, 365]$. Our goal is to find a local window that follows certain optimality properties to be defined below. Specifically, for a specified weight sequence, we conduct a sequential LRT to choose an appropriate bandwidth. Different procedures of estimating seasonality and volatility are studied. Suppose that the object to be approximated is the seasonal variance $\theta(t) = \{\sigma_t^2\}$. A weighted maximum likelihood approach is given by:

$$\begin{aligned}\widetilde{\theta}_k(s) &\stackrel{\text{def}}{=} \arg\max_{\theta\in\Theta} L\{W^k(s), \theta\} \\ &= \arg\min_{\theta\in\Theta} \sum_{t=1}^{365}\sum_{j=0}^{J} \{\log(2\pi\theta)/2 + \widehat{\varepsilon}_{t,j}^2/2\theta\} w(s,t,h_k), \end{aligned} \tag{5.7}$$

with the "localising scheme" $W^k(s) = \{w(s,1,h_k), w(s,2,h_k), \ldots, w(s,365,h_k)\}^\top$, where $w(s,t,h_k) = h_k^{-1} K\{(s-t)/h_k\}$, $k = 1, \ldots, K$, $h_1 < h_2 < h_3 < \ldots < h_K$ the prescribed sequence of bandwidths, and $K(u) = 15/16(1-u^2)^2 \mathbf{I}(|u| \leq 1)$ (quartic kernel).

The explicit solution of (5.7) is given by:

$$\begin{aligned}\widetilde{\theta}_k(s) &= \sum_{t,j} \widehat{\varepsilon}_{t,j}^2 w(s,t,h_k) / \sum_{t,j} w(s,t,h_k) \\ &= \sum_{t} \widehat{\varepsilon}_{t}^2 w(s,t,h_k) / \sum_{t} w(s,t,h_k),\end{aligned}$$

with

$$\widehat{\varepsilon}_t^2 \stackrel{\text{def}}{=} (J+1)^{-1} \sum_{j=0}^{J} \widehat{\varepsilon}_{t,j}^2.$$

From a smoothing perspective we are in a comfortable situation here since the boundary bias is not an issue, as we are dealing with a periodic function $\theta(t) = \theta(t+365)$. We use mirrored observations: assume $h_K < 365/2$, then the observation set, for example for the seasonal volatility, is extended to $\widehat{\varepsilon}_{-364}^2, \widehat{\varepsilon}_{-363}^2, \ldots, \widehat{\varepsilon}_0^2, \widehat{\varepsilon}_1^2, \ldots, \widehat{\varepsilon}_{730}^2$, where

$$\begin{aligned}\widehat{\varepsilon}_t^2 &\stackrel{\text{def}}{=} \widehat{\varepsilon}_{365+t}^2, -364 \leq t \leq 0, \\ \widehat{\varepsilon}_t^2 &\stackrel{\text{def}}{=} \widehat{\varepsilon}_{t-365}^2, 366 \leq t \leq 730.\end{aligned}$$

Since the location s is fixed, we drop s for the simplicity of notation.

For $\ell < k$, the accuracy of the estimation is measured by the fitted likelihood ratio (LR):

$$L(W^\ell, \widetilde{\theta}_\ell, \widetilde{\theta}_k) \stackrel{\text{def}}{=} L(W^\ell, \widetilde{\theta}_\ell) - L(W^\ell, \widetilde{\theta}_k). \tag{5.8}$$

The volatility σ_t or trend Λ_t estimation happens within an exponential family, so LR can be written in closed form, Polzehl & Spokoiny (2006):

$$\begin{aligned} L(W^k, \widetilde{\theta}_k, \theta^*) &\stackrel{\text{def}}{=} N_k \mathcal{K}(\widetilde{\theta}_k, \theta^*) \\ &= -\{\log(\widetilde{\theta}_k/\theta^*) + 1 - \theta^*/\widetilde{\theta}_k\}/2, \end{aligned} \tag{5.9}$$

where $N_k = J \sum_{t=1}^{365} w(s,t,h_k)$ and $\mathcal{K}(\widetilde{\theta}_k, \theta^*)$ is the Kullback-Leibler divergence between two normal distributions with variances $\widetilde{\theta}_k$ and θ^*. Note that (5.9) is the divergence in the volatility case. For trend estimation, it has to be replaced by $(\widetilde{\theta}_k - \theta^*)/(2\sigma^2)$.

The Kullback-Leibler divergence of two distributions with densities $p(x)$ and $q(x)$ is defined as:

$$\mathcal{K}\{p(x), q(x)\} \stackrel{\text{def}}{=} I\!E_{p(.)} \log \frac{p(x)}{q(x)} \tag{5.10}$$

To guarantee the feasibility of the tests, we need moment bounds and confidence sets for LR, which guarantee that the MLE is concentrated in the level set of the likelihood ratio process around the true parameter. For the volatility case, see Polzehl & Spokoiny (2006); for the trend case, see Mercurio & Spokoiny (2004).

Theorem 5.2.1. *[Spokoiny (2009)] Assuming that $\theta(t) = \theta^*$ for any $t \in [1, 365]$, then for $\mathfrak{z} > 0$ and $k \in 1, \ldots, K, r > 0$, denote $\mathrm{P}_{\theta^*}(.)$ as the measure corresponding to (5.7). We obtain:*

$$\mathrm{P}_{\theta^*}\left\{ L(W^k, \widetilde{\theta}_k, \theta^*) > \mathfrak{z} \right\} \leq 2\exp(-\mathfrak{z}) \tag{5.11}$$

and a risk bound for a power loss function:

$$I\!E_{\theta^*} |L(W^k, \widetilde{\theta}_k, \theta^*)|^r \leq \mathfrak{r}_r, \tag{5.12}$$

where $\mathfrak{r}_r = 2r \int_{\mathfrak{z} \geq 0} \mathfrak{z}^{r-1} \exp(-\mathfrak{z}) d\mathfrak{z}$. This polynomial bound applies to all localising schemes W^k simultaneously.

The risk bound (5.12) allows us to define likelihood based confidence sets since together with (5.11) it tells us that the likelihood process is stochastically bounded. Define therefore confidence sets with critical values $\mathfrak{z}_k$ to level α:

$$\mathfrak{E}_{\alpha,k} = \{\theta : L(W^k, \widetilde{\theta}_k, \theta) \leq \mathfrak{z}_k\}. \tag{5.13}$$

Equipped with confidence sets (5.13), we launch the Local Model Selection (LMS) algorithm:

- Fix a point $s \in \{1, 2, \ldots, 365\}$.
- Start with the smallest interval h_1: $\widehat{\theta}_1 = \widetilde{\theta}_1$
- For $k \geq 2$, $\widetilde{\theta}_k$ is accepted and $\widehat{\theta}_k = \widetilde{\theta}_k$ if $\widetilde{\theta}_{k-1}$ was accepted and $\widetilde{\theta}_\ell \in \mathfrak{E}_{\alpha,k}, \forall \ell = 1, \ldots, k-1$, i.e.

  $$L(W^k, \widetilde{\theta}_\ell, \widetilde{\theta}_k) \leq \mathfrak{z}_\ell, \forall \ell = 1, \ldots, k-1.$$

  Otherwise, set $\widehat{\theta}_k = \widehat{\theta}_{k-1}$, where $\widehat{\theta}_k$ is the latest accepted after first k steps.
- Define $\widehat{k}$ as the kth step we stopped, and $\widehat{\theta}_\ell = \widetilde{\theta}_{\widehat{k}}, \ell \geq k$.

The LMS algorithm is illustrated in Figure 5.3. For every estimate $\widetilde{\theta}_k$ the corresponding confidence set is shown. If the horizontal line originating $\widetilde{\theta}_k$ does not cross all the preceding intervals then the selection algorithm terminates.

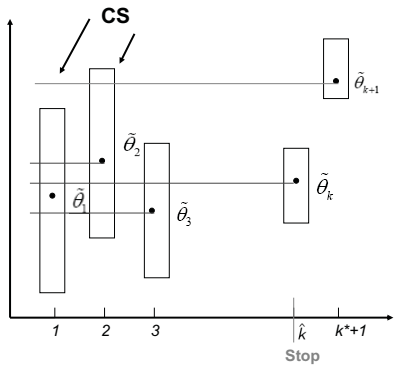


Figure 5.3: Localised model selection (LMS)

A further integrated approach is to consider an iterative algorithm to cope with heteroscedasticity in the corrected residuals after seasonality in mean and variance component varies between estimating the seasonal component and the variance $\theta(t) = \{\Lambda_t, \sigma_t^2\}$. The procedure is:

Step 1. Estimate $\widehat{\beta}$ in an initial Λ_t^0 using a truncated Fourier series or any other deterministic function;

Step 2. For fixed $\widehat{\Lambda}_{s,\nu} = \{\widehat{\Lambda}'_{s,\nu}, \widehat{\Lambda}''_{s,\nu}\}^\top, s = \{1, \ldots, 365\}$ from last step ν, and fixed $\widehat{\beta}$, get $\widehat{\sigma}^2_{s,\nu+1}$ by

$$\begin{aligned}\widehat{\sigma}^2_{s,\nu+1} &= \arg\min_{\sigma^2} \sum_{t=1}^{365}\sum_{j=0}^{J}[\{T_{365j+t} - \widehat{\Lambda}'_{s,\nu} - \widehat{\Lambda}''_{s,\nu}(t-s) \\ &- \sum_{l=1}^{L}\widehat{\beta}_l X_{365j+t-l}\}^2/2\sigma^2 + \log(2\pi\sigma^2)/2]w(s,t,h'_k);\end{aligned}$$

Step 3. For fixed $\widehat{\sigma}^2_{s,\nu+1}$ and $\widehat{\beta}$, we estimate $\widehat{\Lambda}_{s,\nu+1}, s = \{1, \ldots, 365\}$ via another local adaptive procedure:

$$\widehat{\Lambda}_{s,\nu+1} = \arg\min_{\{\Lambda',\Lambda''\}^\top} \sum_{t=1}^{365}\sum_{j=0}^{J}\left\{T_{365j+t} - \Lambda' - \Lambda''(t-s) - \sum_{l=1}^{L}\widehat{\beta}_l X_{365j+t-l}\right\}^2 w(s,t,h'_k)/2\widehat{\sigma}^2_{s,\nu+1},$$

where $\{h'_1, h'_2, h'_3, \ldots, h'_{K'}\}$ is a sequence of bandwidths;

Step 4. Repeat steps 2 and 3 till both $|\widehat{\Lambda}_{t,\nu+1} - \widehat{\Lambda}_{t,\nu}| < \pi_1$ and $|\widehat{\sigma}^2_{t,\nu+1} - \widehat{\sigma}^2_{t,\nu}| < \pi_2$ for some constants π_1 and π_2.

Our empirical implementation suggests that one iteration is enough.

The LMS methods requires critical values $\mathfrak{z}_k$, which define the significance for the LRT statistics $L(W^\ell, \widetilde{\theta}_\ell, \widetilde{\theta}_k)$ or alternatively speaking the length of the confidence interval (see (5.11)) at each step. The critical values are calibrated from the "propagation condition" below which ensures a desired level of type one error. To be more specific, for every step k, define $\widehat{\theta}_k$ as the "survived estimator" after the kth step (if the estimator is not rejected up to step k, then $\widehat{\theta}_k = \widetilde{\theta}_k$, else if the estimator has been rejected at step $l < k$, then $\widehat{\theta}_k = \widetilde{\theta}_l$). Measure the closeness of $\widetilde{\theta}_k$ and $\widehat{\theta}_k$ by:

$$I\!E_{\theta^*}|L(W^k, \widetilde{\theta}_k, \widehat{\theta}_k)|^r \le \alpha\mathfrak{r}_r \tag{5.14}$$

for $k = 1, \ldots, K$ with $\mathfrak{r}_r$ the parametric risk bound in (5.12) and α a control parameter corresponding to the type one error. In fact

$$I\!E_{\theta^*}|L(W^k, \widetilde{\theta}_k, \widehat{\theta}_k)|^r \to \mathrm{P}_{\theta^*}(\widetilde{\theta}_k \neq \widehat{\theta}_k)$$

for $r \to 0$, therefore α can be interpreted as a false alarm probability.

More precisely if step k is accepted as described in Figure 5.3 then $\widetilde{\theta}_k = \widehat{\theta}_k$ and the nonzero loss $I\!E_{\theta^*} L(W^k, \widetilde{\theta}_k, \widehat{\theta}_k)$ can only occur if the estimator has been rejected before or at step k, which under the homogeneous parametric model case, is denoted as "false alarm".

With the "propagation condition" (5.16) below, critical values are constructed.

- Consider first $\mathfrak{z}_1$ and let $\mathfrak{z}_2 = \mathfrak{z}_3 = \ldots = \mathfrak{z}_{K-1} = \infty$. This leads to the estimates $\widehat{\theta}_k(\mathfrak{z}_1)$ and the value $\mathfrak{z}_1$ is selected as the minimal one for which

$$\sup_{\theta^*} I\!E_{\theta^*} |L\{W^k, \widetilde{\theta}_k, \widehat{\theta}_k(\mathfrak{z}_1)\}| \leq \frac{\alpha \mathfrak{r}_r}{K-1}, k = 2, \ldots, K. \tag{5.15}$$

- Suppose $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$ have been fixed, and set $\mathfrak{z}_k = \ldots = \mathfrak{z}_{K-1} = \infty$. With estimate $\widehat{\theta}_m(\mathfrak{z}_1, \ldots, \mathfrak{z}_k)$ for $m = k+1, \ldots, K$. select $\mathfrak{z}_k$ as the minimal value which fulfills

$$\sup_{\theta^*} I\!E_{\theta^*} |L(W^m, \widetilde{\theta}_m, \widehat{\theta}_m(\mathfrak{z}_1, \ldots, \mathfrak{z}_k))|^r \leq \frac{k\alpha \mathfrak{r}_r}{K-1} \tag{5.16}$$

  for $m = k+1, \ldots, K$.

Inequality (5.15) describes the impact of the k critical values to the risk, while the factor $\frac{k\alpha}{K-1}$ in (5.16) ensures that every $\mathfrak{z}_k$ has the same impact. The values of $(\alpha, r, h_1, \ldots, h_K)$ are prespecified hyper-parameters of which robustness and sensitivity issues will be discussed in Section 3. The following theorem provides insight into the form of $\mathfrak{z}_k$.

Theorem 5.2.2. *[Spokoiny (2009)] Suppose that* $0 < h_{k-1}/h_k < 1$ *and* $\theta(t) = \theta^*$ *for all* $t \in [0, 365]$. *An upper bound for the critical values* $\mathfrak{z}_k$ *is given by:*

$$\mathfrak{z}_k = a_0 \log K + 2 \log(n h_k/\alpha) + 2r \log(h_K/h_k)$$

where $a_0 > 0$ *is a constant.*

A risk bound for a global model ($\theta(t) = \theta^*$) has been given in (5.14). This may now be extended to the nonparametric setting via the "Small Modeling Bias (SMB)" condition:

$$\Delta(\theta) \stackrel{\text{def}}{=} \sum_{t=1}^{365} \mathcal{K}(\theta_t, \theta) \mathbf{I}\{w(s, t, h_k) > 0\} \leq \Delta, \forall k < k^*, \tag{5.17}$$

where k^* is the maximum k satisfying (5.17), also called "oracle".

The estimation risk for the function $\theta(t)$ is described for $k \leq k^*$ by the "propagation" property:

$$I\!E_{\theta(\cdot)} \log\{1 + |L(W^k, \widetilde{\theta}_k, \widehat{\theta}_k)|^r / \mathfrak{r}_r\} \leq \Delta + \alpha. \tag{5.18}$$

An estimate for k^* is desired. The adaptive estimate $\widehat{\theta}_{\widehat{k}}$ will in fact enjoy this property as we show below. The estimate $\widehat{\theta}_{\widehat{k}}$ behaves similarly to the oracle estimate $\widetilde{\theta}_{k^*}$ since it is "stable" in the sense that even if the described selection scheme overshoots k^*, the resulting estimate $\widehat{\theta}_{\widehat{k}}$ is still close to the oracle $\widetilde{\theta}_{k^*}$. This may be expressed as that the attained quality of estimation during "propagation" is not lost at further steps:

$$L(W^{k^*}, \widetilde{\theta}_{k^*}, \widehat{\theta}_{\widehat{k}})\, \mathbf{I}\{\widehat{k} > k^*\} \leq \mathfrak{z}_{k^*}$$

A combination of the propagation and stability property then leads to the "oracle" property:

$$\begin{aligned} I\!E_{\theta(\cdot)} \log\left\{1 + \frac{|L(W^{k^*}, \widetilde{\theta}_{k^*}, \theta)|^r}{\mathfrak{r}_r}\right\} &\leq \Delta + 1, \\ I\!E_{\theta(\cdot)} \log\left\{1 + \frac{|L(W^{k^*}, \widetilde{\theta}_{k^*}, \widehat{\theta}_{\widehat{k}})|^r}{\mathfrak{r}_r}\right\} &\leq \Delta + \alpha + \log\left\{1 + \frac{\mathfrak{z}_{k^*}}{\mathfrak{r}_r}\right\}, \end{aligned}$$

for $\theta \in \Theta$ with $\Delta(W^k, \theta) \leq \Delta$ and $k \leq k^*$. This means that the risk of estimating adaptively is composed into three parts: the SMB, the false alarm rate and a small term corresponding to the overshooting risk.

5.3 Empirical analysis

We conduct an empirical analysis of temperature patterns over different cities (Figure 5.4). The data set contains daily average temperatures for different cities in Europe, Asia and US: Atlanta, Beijing, Berlin, Essen, Houston, Kaoshiung, New York, Osaka, Portland, Taipei, Tokyo. The summary of the data and characteristics can be seen in Table 5.1.

We first check seasonality, intertemporal correlation and seasonal variation. Table 5.2 provides the coefficients of the Fourier truncated seasonal function (5.1) for some cities for different time periods. The coefficient a can be seen as the average temperature, the coefficient b as an indicator for global warming. The latter coefficients are stable even when the estimation is done in a window length of 10 years. In the sense of capturing volatility peak seasons, the left panel of Figure 5.5 visualizes the power of capturing volatility peak seasons by the seasonal

Figure 5.4: Map of locations where temperature are collected.

City	Period	ADF	KPSS	$AR(3)$			$CAR(3)$		
		$\widehat{\tau}$	$\widehat{k}$	β_1	β_2	β_3	α_1	α_2	α_3
Atlanta	19480101-20081204	-55.55+	0.21***	0.96	-0.38	0.13	2.03	1.46	0.28
Beijing	19730101-20090831	-30.75+	0.16***	0.72	-0.07	0.05	2.27	1.63	0.29
Berlin	19480101-20080527	-40.94+	0.13**	0.91	-0.20	0.07	2.08	1.37	0.20
Essen	19700101-20090731	-23.87+	0.11*	0.93	-0.21	0.11	2.06	1.34	0.16
Houston	19700101-20081204	-38.17+	0.05*	0.90	-0.39	0.15	2.09	1.57	0.33
Kaohsiung	19730101-20091210	-37.96+	0.05*	0.73	-0.08	0.04	2.26	1.60	0.29
New York	19490101-20081204	-56.88+	0.08*	0.76	-0.23	0.11	2.23	1.69	0.34
Osaka	19730101-20090604	-18.65+	0.09*	0.73	-0.14	0.06	2.26	1.68	0.34
Portland	19480101-20081204	-45.13+	0.05*	0.86	-0.22	0.08	2.13	1.48	0.26
Taipei	19920101-20090806	-32.82+	0.09*	0.79	-0.22	0.06	2.20	1.63	0.36
Tokyo	19730101-20090831	-25.93+	0.06*	0.64	-0.07	0.06	2.35	1.79	0.37

Table 5.1: ADF and KPSS-Statistics, coefficients of the autoregressive process $AR(3)$ and continuous autoregressive model $CAR(3)$ model for the detrended daily average temperatures time series for different cities. +0.01 critical values, * 0.1 critical value, **0.05 critical value, ***0.01 critical value.

City	Period	$\widehat{a}$	$\widehat{b}$	$\widehat{c}_1$	$\widehat{d}_1$	$\widehat{c}_2$	$\widehat{d}_2$	$\widehat{c}_3$	$\widehat{d}_3$
Berlin	(19480101-20080527)	9.2173	0.0000	9.8932	-157.9123	0.2247	261.2850	0.1591	-127.7303
	(19730101-20080527)	9.3050	0.0001	10.0070	-161.2493	0.4601	-66.0530	-0.3723	-416.4776
	(19730101-20080527)	9.3050	0.0001	10.0070	-161.2493	0.4601	-66.0530	-0.3723	-416.4776
	(19830101-20080527)	9.4581	0.0001	10.0969	-161.7129	0.5205	-51.9929	0.3734	42.0874
	(19930101-20080527)	9.5923	0.0002	10.1995	-162.9774	0.6564	-37.1548	0.4241	41.9970
	(20030101-20080527)	9.6948	0.0007	10.1954	-162.3343	0.5554	-43.2293	0.3269	1.5998
Kaohsiung	(19730101-20081231)	24.2289	0.0001	0.9157	-145.6337	-4.0603	-78.1426	-1.0505	10.6041
	(19730101-19821231)	24.4413	0.0001	2.1112	-129.1218	-3.3887	-91.1782	-0.8733	20.0342
	(19830101-19921231)	25.0616	0.0003	2.0181	-135.0527	-2.8400	-89.3952	-1.0128	20.4010
	(19930101-20021231)	25.3227	0.0003	3.9154	-165.7407	-0.7405	-51.4230	-1.1056	19.7340
New-York	(19490101-20081204)	53.1473	0.0001	18.6810	-143.4051	-3.3872	271.5072	-0.4203	-16.3125
	(19730101-20081204)	53.6992	0.0001	18.0092	-148.4124	-3.5236	279.6876	-0.4756	-21.8090
	(19730101-19821204)	53.6037	-0.0000	17.7446	-155.2453	-3.7769	289.7932	-0.8326	-4.2257
	(19830101-19921204)	54.8740	-0.0003	17.6924	-152.7461	-3.4245	284.6412	-0.4933	-218.9204
	(19930101-20021204)	53.8050	0.0003	17.6942	-153.3997	-3.4246	285.7958	0.5753	-315.2792
	(20030101-20081204)	52.9177	0.0012	17.8425	-151.2977	-3.8837	287.2022	-0.1290	-216.7298
Tokyo	(19730101-20081231)	15.7415	0.0001	8.9171	-162.3055	-2.5521	-7.8982	-0.7155	-15.0956
	(19730101-19821231)	15.8109	0.0001	9.2855	-162.6268	-1.9157	-16.4305	-0.5907	-13.4789
	(19830101-19921231)	15.4391	0.0004	9.4022	-162.5191	-2.0254	-4.8526	-0.8139	-19.4540
	(19930101-20021231)	16.4284	0.0001	8.8176	-162.2136	-2.1893	-17.7745	-0.7846	-22.2583
	(20030101-20081231)	16.4567	0.0001	8.5504	-162.0298	-2.3157	-18.3324	-0.6843	-16.5381

Table 5.2: Seasonality estimates $\widehat{\Lambda}_t$ of daily average temperatures in Asia. All coefficients are nonzero at 1% significance level. Data source: Bloomberg.

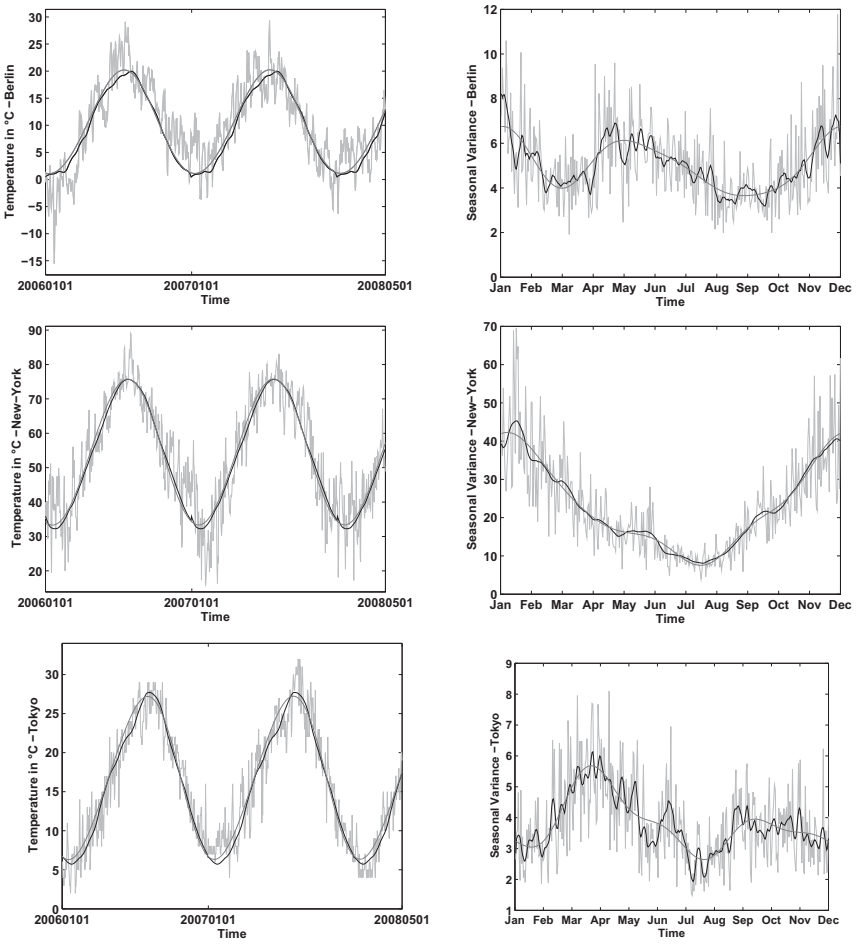


Figure 5.5: The empirical (black line), the Fourier truncated (dotted gray line) and the the local linear (gray line) seasonal mean (left panel) and variance component (right panel) using Quartic kernel and bandwidth $h = 4.49$.

City	Corrected residuals with Fourier $\frac{\widehat{\varepsilon}_t}{\widehat{\sigma}_{t,FTSG}}$					Corrected residuas with Local smoother $\frac{\widehat{\varepsilon}_t}{\widehat{\sigma}_{t,LLR}}$				
	JB	Kurtosis	Skewness	KS	AD	JB	Kurtosis	Skewness	KS	AD
Berlin	304.77	3.54	-0.08	0.01	7.65	279.06	3.52	-0.08	0.01	7.29
New-York	403.39	3.47	-0.23	0.02	23.22	375.50	3.45	-0.228	0.02	21.74
Kaohsiung	2753.00	4.68	-0.71	0.06	79.93	2252.50	4.52	-0.64	0.06	79.18
Tokyo	133.26	3.44	-0.10	0.02	8.06	148.08	3.44	-0.13	0.02	10.31

Table 5.3: Skewness, kurtosis, Jarque Bera (JB), Kolmogorov Smirnov (KS) and Anderson Darling (AD) test statistics (365 days) of corrected residuals.

local smoother (5.4) using the quartic kernel over the estimates modeled under Fourier truncated series (5.1).

After removing the local linear seasonal mean (5.4) from the daily average temperatures ($X_t = T_t - \Lambda_{t,LNN}$), we check that X_t is a stationary process with the Augmented Dickey-Fuller (ADF) and the KPSS tests. The analysis of the partial autocorrelations and Akaike's Information criterion (AIC) suggest that a simple $AR(3)$ model fits the temperature evolution well. Table 5.1 presents the results of the stationarity tests as well as the coefficients of the fitted $AR(3)$. The empirical seasonal variation (square residuals after seasonal and intertemporal fitting), the seasonal variation curves (5.3) and (5.5) are displayed on the right panel in Figure 5.5, while the descriptive statistics for the residuals after correcting by seasonality are given in Table 5.3. Both seasonal volatility estimators lead to heavy tail distributions of corrected residuals and negative skewness.

The adjustment in the smoothing parameter h will provide the localisation in time. The bandwidth sequences are selected from four candidates: $(3, 5, 7, 9, 11, 13, 15)$, $(3, 5, 8, 12, 17, 23, 30)$, $(5, 7, 10, 14, 19, 25, 32)$, $(7, 9, 11, 14, 17, 10, 24)$. The candidates are chosen according to the lowest Anderson Darling statistic. The best candidate for bandwidth sequence is that one that yields a residual distribution close to normality. Smoothing the bandwidths selected at discrete points, gives yet another adaptive estimator.

The critical values (CV) as calibrated from (5.15) and (5.16) are given in Figure 5.6. The left side provides CVs simulated from a sample of 10^3 observations for a quartic kernel for both mean and volatility with $\theta^* = 1$, $r = 0.5$ and different values of significance level α. The CVs for different bandwidth sequences are displayed in the right side of Figure 5.6. The CVs, as one observes, are insensitive to the choice of r and α.

A one year short period is considered in the first place for demonstration purpose, while later we show how the results change with different time length periods. Figures 5.7, 5.8, 5.9 and 5.10 present general results for different cities under dif-

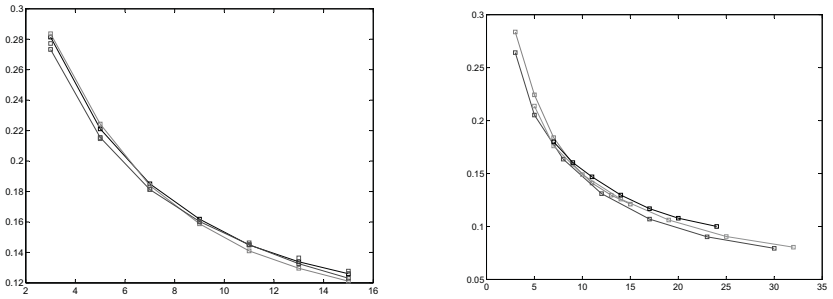
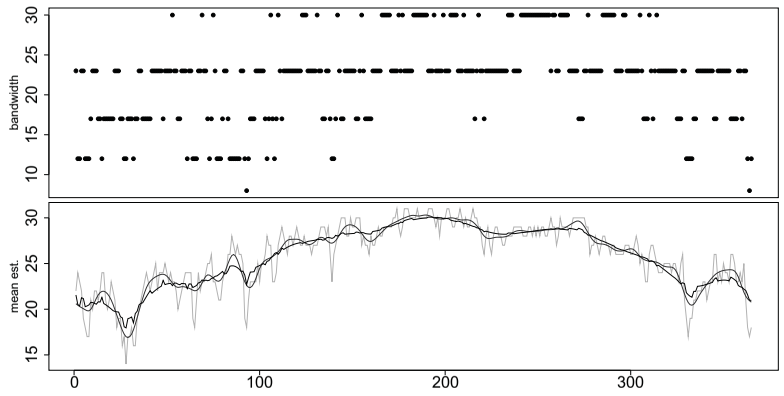
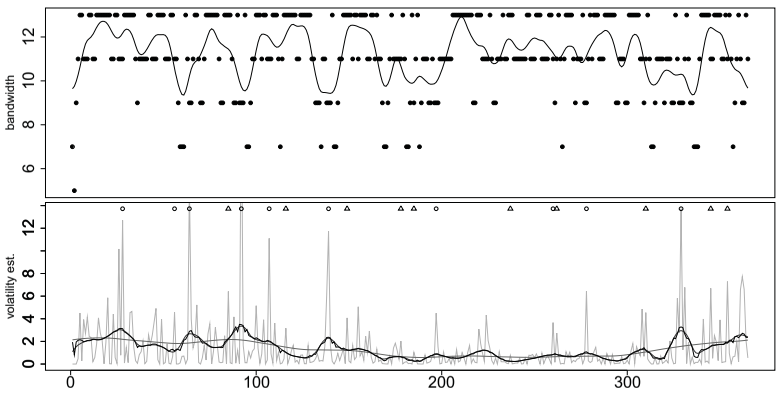


Figure 5.6: Simulated CV for likelihood of seasonal volatility (5.7) with $\theta^* = 1$, $r = 0.5$, $MC = 5000$ with $\alpha = 0.3$ (gray dotted line), 0.5 (black dotted line), 0.8 (dark gray dotted line) (left), with different bandwidth sequences (right).

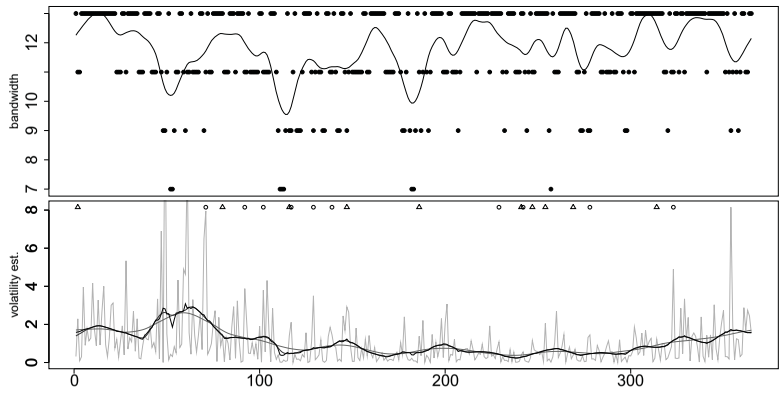
ferent adaptive localising schemes for seasonal mean (Me) and seasonal volatility (Vo): with fixed bandwidth curve (fi), adaptive bandwidth curve (ad) and adaptive smoothed bandwidth (ads) for different time intervals. The seasonal mean is estimated jointly over the years, using $\alpha = 0.3$ and power level $r = 0.5$. The upper panel of each volatility plot on Figures 5.7-5.10 shows the sequence of bandwidths and the smoothed bandwidth; the bottom panel displays the variance estimation with fixed bandwidth (dashed line), smoothed adaptive bandwidth (dotted line) and adaptive bandwidth (dot-dashed line). In all countries, one observes significant differences between the estimates. When smoothing the discrete bandwidths over time, the estimated variance curves are smoother. In particular, in cities like Kaohsiung and New York, one observes more variation of the seasonal variance curves during peak seasons (winter and summer times). The triangles and circles in the bottom panel of each volatility plot helps us to trace the source of non-normality over time, since they corresponds to 10 dots of the upper and lower tails of the QQ-plots of square residuals respectively (see Figure 5.11 for Berlin results). Left top plots of Figures 5.7- 5.10 show the mean case. Different from the seasonal variance function, we do not observe a big variation of smoothness in the mean function. One can see that in all cities, the bandwidths are varying over the yearly cycle with a slight degree of non homogeneity for Kaoshiung.



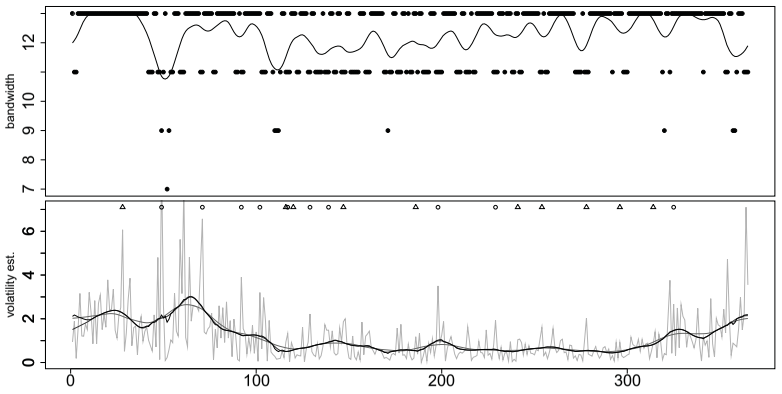
(a) Mean, 2008



(b) Volatility, 2008

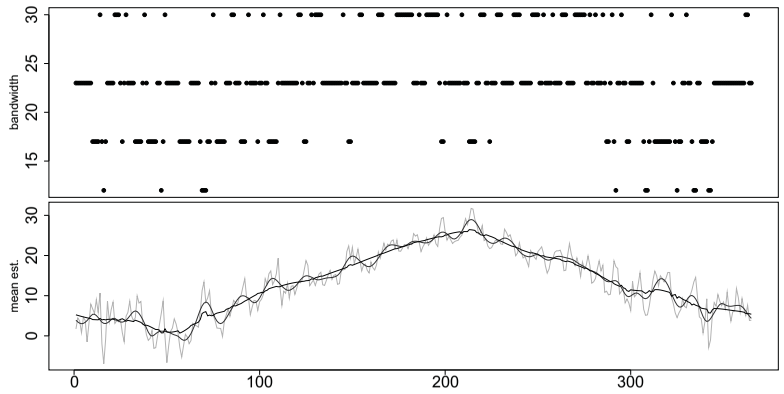


(c) Volatility, 2006-2008

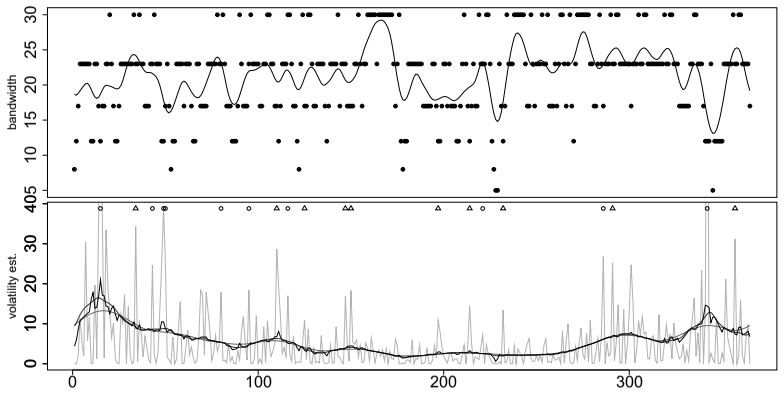


(d) Volatility, 2004-2008

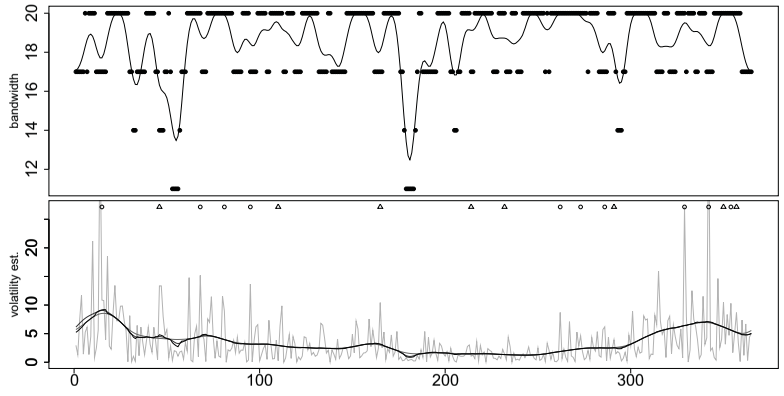
Figure 5.7: Estimation of mean and variance for Kaohsiung. In each figure sequence (also smoothed for volatility)



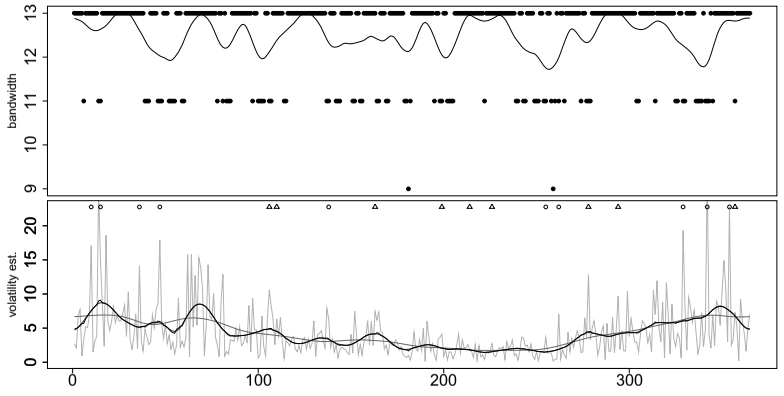
(a) Mean, 2007



(b) Volatility, 2007

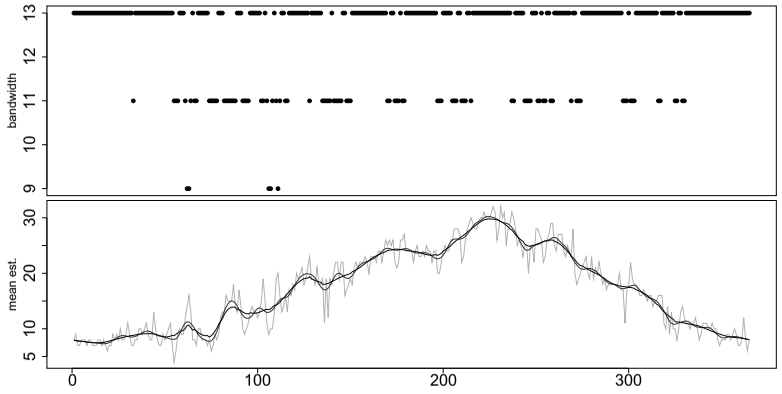


(c) Volatility, 2005-2007

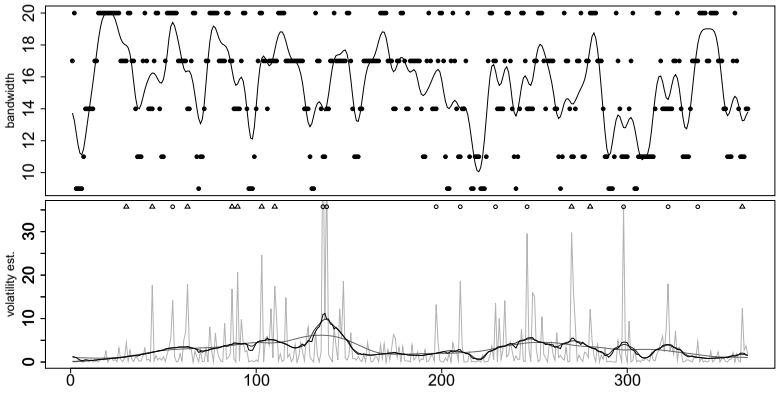


(d) Volatility, 2003-2007

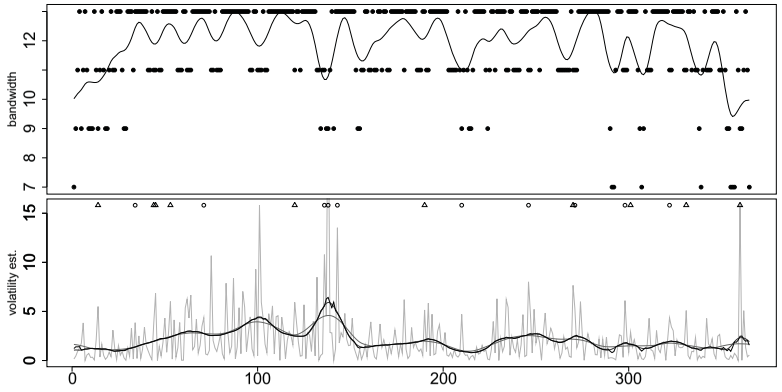
Figure 5.8: Estimation of mean and variance for New-York. In each figure sequence (also smoothed for volatility) of bandwidths (upper panel), nonparametric function estimation (solid grey line), with fixed bandwidth (dashed line), adaptive bandwidth (dot-dashed line) and smoothed adaptive bandwidth (dotted line) (bottom panel of each



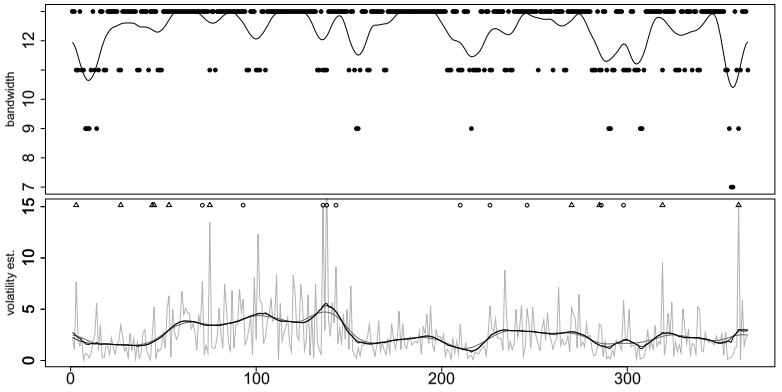
(a) Mean, 2008



(b) Volatility, 2008

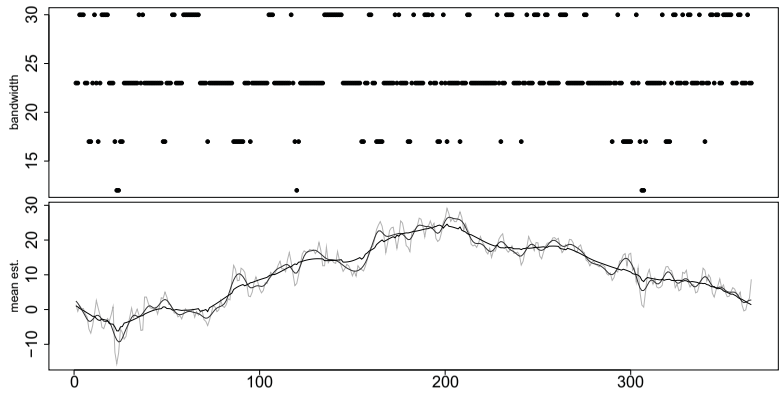


(c) Volatility, 2006-2008

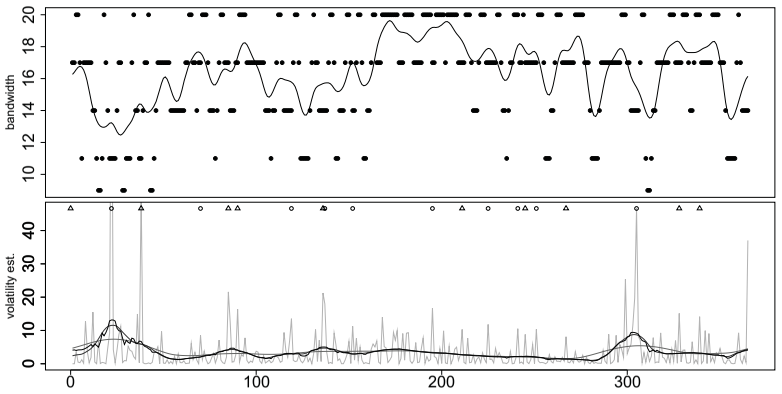


(d) Volatility, 2004-2008

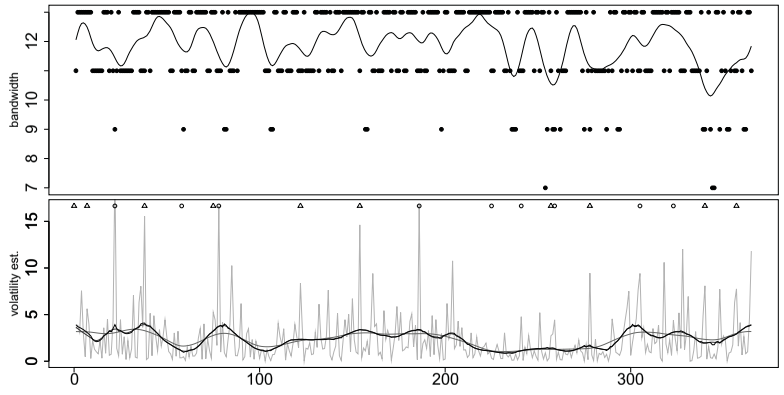
Figure 5.9: Estimation of mean and variance for Tokyo. In each figure sequence (also smoothed for volatility) of



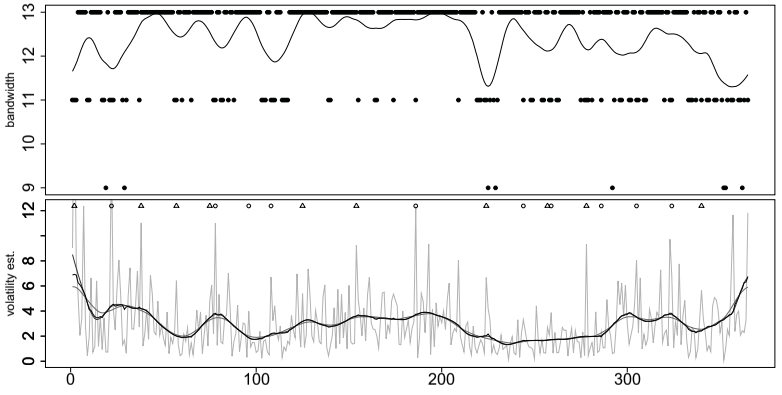
(a) Mean, 2007



(b) Volatility, 2007



(c) Volatility, 2005-2007



(d) Volatility, 2003-2007

Figure 5.10: Estimation of mean and variance for Berlin. In each figure sequence (also smoothed for volatility) of bandwidths (upper panel), nonparametric function estimation (solid grey line), with fixed bandwidth (dashed line), adaptive bandwidth (dot-dashed line) and smoothed adaptive bandwidth (dotted line) (bottom panel of each figure).

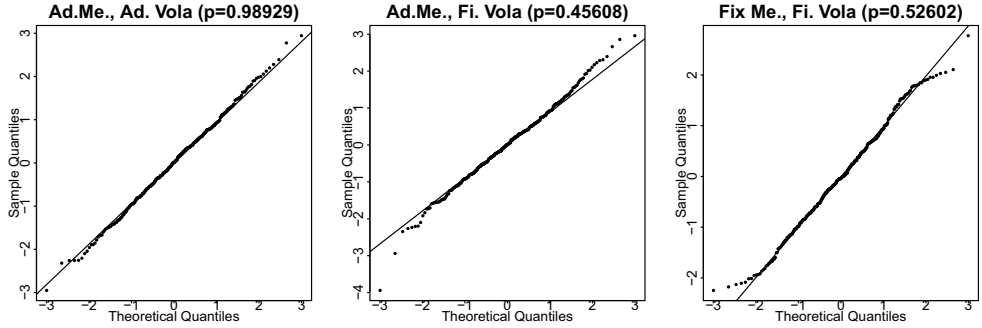
An approach to cope with the non normality brought in by more observations is to estimate mean functions year by year (SeMe), and then aggregate the residuals for variance estimation. We therefore estimate the joint/separate seasonal mean (JoMe/SeMe) and seasonal variance (Vo) curves with fixed bandwidth curve (fi), adaptive bandwidth curve (ad) and adaptive smoothed bandwidth (ads). Table 5.5 and Table 5.6 show the p-values for normality tests. Volatility plots on the Figures 5.7-5.10 displays the behavior of the variance function estimation when the period length changes. The average over years acts as a smoother when we consider more years. The estimated $AR(L)$ parameters for different cities using joint/separate mean (JoMe/SeMe) with different bandwidth curves are illustrated in Table 5.4. The results again show that an $AR(3)$ fits well the stylised facts of temperature.

The p-values of normality test statistics (Kolmogorov Smirnov KS, Jarques-Bera JB, Anderson Darling AD) of corrected residuals (after seasonal mean and volatility) for different cities under varying localising schemes are displayed in Table 5.5 and Table 5.6. The results are compared for different periods (3 years, 4 years, 5 years). The longer the period, the smaller the p-value of normality and therefore the more likely to reject the normality assumption. The standardised residuals are closer to normality (Berlin and New York) or at the same level (Kaoshiung and Tokyo) overall. The approach shows stability over more years. The p-values for adaptive estimates, over all cities, are generally larger than those for fixed bandwidth estimates. We observe that in US cities the risk factor show a better Gaussian pattern compared to other cities. With smoothed bandwidth, there are a slightly improvements in some cases. In most of the cases, specially in cities at sea level, the correction by adaptive models outperforms the classical method.

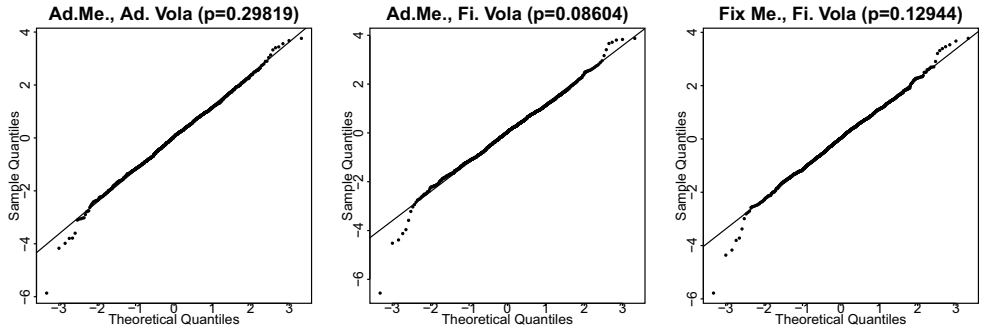
We tackle the problem of loosing information when considering estimates at individual level or averaging mean functions over time, with a refined approach that considers the minimum variance between the aggregation of yearly local mean function estimates and an optimal local estimate θ^o. Once the sets of local mean functions have been identified, the aggregated local function can be defined as the weighted average of all the observations in a given time set. Formally, if $\widehat{\theta^j}(t)$ is the localised observation at time t of year j, the aggregated local function is given by:

$$\widehat{\theta}_\omega(t) = \sum_{j=1}^{J} \omega_j \widehat{\theta^j}(t). \tag{5.19}$$

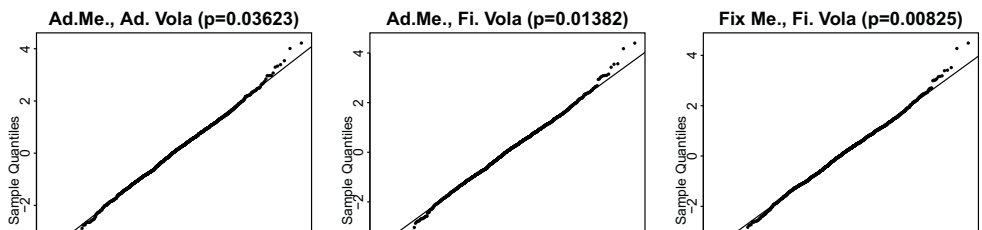
With this aggregation step across J, we give the same weight to all observations, even to observations that were unimportant at the yearly level. Then a reasonable



(a) Berlin 1 year (2007)



(b) Berlin 3 years (2004-2007)



City	Method	Period	Mean	α_1	α_2	α_3	mean (AR)
Berlin	JoMe	5 years	ad	0.9970	-0.2923	0.0969	6.29e-03
			fi	0.9776	-0.2899	0.1129	4.80e-16
	SeMe	1 year	ad	0.8291	-0.2758	0.0000	-7.13e-03
			fi	0.3091	-0.3294	-0.2674	-9.65e-16
		2 years	ad	0.8153	-0.2574	-0.0578	-9.29e-03
			fi	0.3553	-0.3318	-0.1959	-5.46e-16
		3 years	ad	0.8481	-0.2793	0.0000	-2.21e-02
			fi	0.3564	-0.3333	-0.1769	-6.88e-16
		4 years	ad	0.8009	-0.2553	0.0000	-5.36e-04
			fi	0.3026	-0.3312	-0.1751	-7.49e-16
		5 years	ad	0.8357	-0.2570	0.0000	5.49e-03
			fi	0.3333	-0.3413	-0.1654	-6.63e-16
Tokyo	JoMe	5 years	ad	0.5985	-0.1006	0.0697	-1.49e-02
			fi	0.5760	-0.1057	0.0716	1.68e-16
	SeMe	1 year	ad	0.3191	-0.0570	-0.1939	2.17e-03
			fi	0.1510	-0.1538	-0.2985	-3.50e-16
		2 years	ad	0.4690	-0.0736	-0.0929	5.54e-03
			fi	0.2239	-0.1785	-0.2459	-1.03e-15
		3 years	ad	0.4486	-0.1355	-0.0628	-6.14e-03
			fi	0.2660	-0.2113	-0.1842	-1.14e-15
		4 years	ad	0.4752	-0.1315	-0.0445	-9.97e-04
			fi	0.2719	-0.2116	-0.1701	-1.21e-15
		5 years	ad	0.4334	-0.1562	-0.0578	-2.63e-03
			fi	0.2546	-0.2306	-0.1704	-1.02e-15
NewYork	JoMe	5 years	ad	0.7333	-0.1956	0.1202	1.66e-03
			fi	0.7128	-0.1966	0.1375	-1.26e-16
	SeMe	1 year	ad	0.6467	-0.1745	0.0000	-2.49e-03
			fi	0.3440	-0.2773	-0.1180	4.90e-16
		2 years	ad	0.5994	-0.2111	0.0000	6.43e-04
			fi	0.2809	-0.3114	-0.1188	3.87e-16
		3 years	ad	0.5581	-0.2355	0.0000	2.00e-04
			fi	0.2888	-0.3208	-0.1319	1.82e-16
		4 years	ad	0.5447	-0.2404	0.0000	-1.29e-03
			fi	0.3039	-0.3248	-0.1186	1.24e-16
		5 years	ad	0.5425	-0.2353	0.0000	-1.47e-03
			fi	0.2925	-0.3073	-0.1509	2.13e-16
Kaohsiung	JoMe	5 years	ad	0.7818	-0.1416	0.0000	2.37e-04
			fi	0.7661	-0.1320	0.0000	7.12e-16
	SeMe	1 year	ad	0.6747	-0.0948	-0.0920	4.87e-02
			fi	0.4719	-0.1740	-0.2126	6.22e-16
		2 years	ad	0.6178	-0.0854	-0.1348	3.92e-02
			fi	0.4767	-0.1596	-0.2182	-2.43e-17
		3 years	ad	0.6740	-0.1628	-0.1149	4.01e-02
			fi	0.4751	-0.2249	-0.2131	-2.20e-16
		4 years	ad	0.6387	-0.1250	-0.1164	3.39e-02
			fi	0.4505	-0.1916	-0.2030	1.10e-16
		5 years	ad	0.6283	-0.1286	-0.0991	3.04e-02
			fi	0.4262	-0.1967	-0.1965	-1.93e-16

Table 5.4: $AR(L)$ parameters for Berlin (20020101-20071201), Tokyo (20030101-20081201), New-York (20030101-20081201) and Kaohsiung (20030101-20081201) using joint/separate mean (JoMe/SeMe) with fixed bandwidth curve (fi), adaptive bandwidth curve (ad), adaptive smoothed bandwidth (ads) seasonal mean/volatility (Me/Vo) curve.

	Method	p-Values (1year)			p-Values (2years)			p-Values (3 years)			p-Values (4years)			p-Values(5 years)		
		KS	JB	AD	KS	JB	AD	KS	JB	AD	KS	JB	AD	KS	JB	AD
Berlin	JoMe adMe fiVo	0.9004	0.0265	0.4560	1.6e-04	2.6e-14	0.0929	0.0142	3.8e-14	0.0860	0.1026	8.8e-16	0.0116	0.1339	2.9e-14	0.0138
	JoMe adMe adVo	0.9917	0.9907	0.9892	9.9e-05	3.0e-02	0.4161	0.0101	4.7e-05	0.2981	0.0473	1.1e-08	0.0299	0.1417	1.0e-08	0.0362
	JoMe adMe adsVo	0.9797	0.9907	0.9892	1.1e-04	3.0e-02	0.4161	0.0113	4.7e-05	0.2981	0.0572	1.1e-08	0.0299	0.1173	1.0e-08	0.0362
	JoMe fiMe fiVo	0.8205	0.3713	0.5260	3.2e-03	1.3e-10	0.0609	0.0888	4.0e-08	0.1294	0.0475	4.5e-13	0.0034	0.3376	7.3e-14	0.0082
	JoMe fiMe adVo	0.7985	0.3369	0.6255	5.4e-04	9.6e-02	0.1420	0.0342	4.2e-03	0.2582	0.0156	1.6e-07	0.0112	0.2689	1.9e-07	0.0235
	JoMe fiMe adsVo	0.7602	0.3369	0.6255	4.4e-04	9.6e-02	0.1420	0.0336	4.2e-03	0.2582	0.0211	1.6e-07	0.0112	0.2370	1.9e-07	0.0235
	SeMe adMe fiVo				8.3e-04	0.0e+00	0.0378	0.0114	1.9e-10	0.0584	0.1326	1.2e-06	0.1021	0.1691	4.1e-08	0.1250
	SeMe adaMe adVo				9.0e-05	3.3e-02	0.6452	0.0079	2.9e-03	0.2060	0.0733	4.6e-03	0.2429	0.1800	2.3e-05	0.2274
	SeMe adaMe adsVo				7.9e-05	3.3e-02	0.6452	0.0055	2.9e-03	0.2060	0.0672	4.6e-03	0.2429	0.1840	2.3e-05	0.2274
	SeMe fiMe fiVo				5.6e-04	2.8e-03	0.0410	0.0059	1.6e-05	0.0110	0.0917	4.0e-05	0.0121	0.0969	6.8e-05	0.0338
	SeMe fiMe adVo				5.7e-04	6.8e-02	0.1280	0.0061	9.5e-03	0.0624	0.0980	6.9e-03	0.0548	0.0904	1.6e-03	0.0708
	SeMe fiMe adsVo				4.7e-04	6.8e-02	0.1280	0.0060	9.5e-03	0.0624	0.1061	6.9e-03	0.0548	0.0860	1.6e-03	0.0708
	SeMe Locave				9.8e-01	1.9e-01	0.8237	0.8473	3.0e-04	0.1181	0.5131	1.9e-03	0.1045	0.4791	1.7e-05	0.1039
	SeMe Locsep				9.8e-01	1.9e-01	0.8237	0.8475	3.0e-04	0.1181	0.5127	1.9e-03	0.1045	0.4803	1.7e-05	0.1041
	SeMe Locmax				9.7e-01	1.6e-01	0.8257	0.8102	1.8e-04	0.1194	0.8727	3.6e-04	0.1555	0.5898	2.5e-05	0.1771
Kaohsiung	JoMe adMe fiVo	0.1015	9.3e-07	4.7e-05	1.2e-04	0.0e+00	6.3e-09	0.0014	0.0e+00	4.8e-15	0.0011	0.0e+00	2.5e-19	5.0e-05	0.0e+00	2.0e-22
	JoMe adMe adVo	0.3454	1.7e-02	5.5e-03	5.6e-05	7.7e-09	6.8e-06	0.0012	0.0e+00	5.4e-13	0.0007	0.0e+00	1.3e-17	3.8e-05	0.0e+00	5.3e-21
	JoMe adMe adsVo	0.3401	1.7e-02	5.5e-03	7.0e-05	7.7e-09	6.8e-06	0.0008	0.0e+00	5.4e-13	0.0010	0.0e+00	1.3e-17	3.8e-05	0.0e+00	5.3e-21
	JoMe fiMe fiVo	0.4050	6.7e-04	8.5e-04	3.2e-05	0.0e+00	3.6e-09	0.0023	0.0e+00	1.4e-14	0.0038	0.0e+00	8.7e-19	5.8e-05	0.0e+00	2.3e-21
	JoMe fiMe adVo	0.3595	8.5e-03	1.9e-03	2.5e-05	5.3e-11	1.5e-06	0.0011	0.0e+00	5.9e-13	0.0038	0.0e+00	1.2e-17	6.8e-05	0.0e+00	2.0e-20
	JoMe fiMe adsVo	0.4203	8.5e-03	1.9e-03	5.1e-05	5.3e-11	1.5e-06	0.0020	0.0e+00	5.9e-13	0.0038	0.0e+00	1.2e-17	6.8e-05	0.0e+00	2.0e-20
	SeMe adMe fiVo				1.4e-05	1.8e-11	2.4e-07	0.0007	0.0e+00	9.5e-11	0.0023	0.0e+00	7.8e-14	2.2e-03	0.0e+00	4.6e-16
	SeMe adMe adVo				1.2e-06	1.5e-06	1.4e-05	0.0003	3.9e-15	3.7e-09	0.0012	0.0e+00	3.4e-12	1.8e-03	0.0e+00	1.1e-14
	SeMe adMe adsVo				3.0e-06	1.5e-06	1.4e-05	0.0004	3.9e-15	3.7e-09	0.0011	0.0e+00	3.4e-12	1.9e-03	0.0e+00	1.1e-14
	SeMe fiMe fiVo				5.8e-06	2.0e-06	3.1e-04	0.0011	1.2e-07	2.2e-05	0.0196	1.0e-13	9.7e-09	1.2e-02	2.1e-13	3.1e-09
	SeMe fiMe adVo				3.1e-06	5.7e-04	6.8e-03	0.0006	5.5e-07	4.8e-05	0.0200	2.5e-12	4.1e-08	1.6e-02	3.3e-15	1.4e-09
	SeMe fiMe adsVo				4.0e-06	5.7e-04	6.8e-03	0.0011	5.5e-07	4.8e-05	0.0203	2.5e-12	4.1e-08	1.4e-02	3.3e-15	1.4e-09
	SeMe Locave				8.0e-02	4.6e-06	3.0e-06	0.0241	6.9e-12	5.5e-09	0.0126	0.0e+00	1.9e-12	1.2e-03	0.0e+00	1.1e-14
	SeMe Locsep				8.0e-02	4.6e-06	3.0e-06	0.0241	6.8e-12	5.5e-09	0.0126	0.0e+00	1.9e-12	1.2e-03	0.0e+00	1.1e-14
	SeMe Locmax				8.3e-02	5.7e-08	1.0e-05	0.0333	5.2e-14	4.7e-09	0.0144	0.0e+00	2.2e-12	2.7e-03	0.0e+00	1.6e-14

Table 5.5: *p*s-values of Jarque Bera (JB), Kolmogorov Smirnov (KS) and Anderson Darling (AD) test statistics for Berlin (20020101-20071201) & Kaohsiung (20020101-20071201) corrected residuals under different adaptive localizing schemes: for joint/separate mean (JoMe/SeMe) with fixed bandwidth curve (fi), adaptive bandwidth curve (ad), adaptive smoothed bandwidth (ads) seasonal mean/volatility (Me/Vo) curve.

	Method	p-Values (1year)			p-Values (2years)			p-Values (3 years)			p-Values (4years)			p-Values(5 years)		
		KS	JB	AD	KS	JB	AD	KS	JB	AD	KS	JB	AD	KS	JB	AD
New-York	JoMe adMe fiVo	0.8677	0.6282	0.3117	6.3e-04	0.0406	0.1975	4.1e-03	0.0097	0.0350	0.0267	0.0080	0.0190	0.0237	0.0008	0.0012
	JoMe adMe adVo	0.9149	0.1085	0.1066	2.6e-04	0.3108	0.3198	1.9e-03	0.0503	0.1068	0.0173	0.0598	0.1153	0.0154	0.0203	0.0167
	JoMe adMe adsVo	0.9512	0.1085	0.1066	2.3e-04	0.3108	0.3198	2.6e-03	0.0503	0.1068	0.0150	0.0598	0.1153	0.0172	0.0203	0.0167
	JoMe fiMe fiVo	0.6061	0.2022	0.0200	4.0e-04	0.0130	0.1118	1.6e-02	0.0027	0.0141	0.0966	0.0050	0.0238	0.0523	0.0006	0.0043
	JoMe fiMe adVo	0.5892	0.0210	0.0039	2.4e-04	0.2331	0.2557	1.7e-02	0.0202	0.0433	0.0461	0.0560	0.1313	0.0457	0.0159	0.0315
	JoMe fiMe adsVo	0.6095	0.0210	0.0039	4.8e-04	0.2331	0.2557	1.7e-02	0.0202	0.0433	0.0562	0.0560	0.1313	0.0521	0.0159	0.0315
	SeMe adMe fiVo				2.7e-06	0.0556	0.3171	1.3e-04	0.5894	0.2341	0.0013	0.3321	0.1727	0.0164	0.1087	0.0299
	SeMe adMe adVo				3.9e-06	0.3266	0.7074	1.5e-04	0.7079	0.3133	0.0009	0.4963	0.2271	0.0076	0.4263	0.0950
	SeMe adMe adsVo				7.2e-06	0.3266	0.7074	1.7e-04	0.7079	0.3133	0.0010	0.4963	0.2271	0.0069	0.4263	0.0950
	SeMe fiMe fiVo				1.4e-06	0.0384	0.0292	1.9e-04	0.5605	0.0598	0.0012	0.1405	0.0197	0.0235	0.0025	0.0073
	SeMe fiMe adVo				2.1e-07	0.0360	0.0259	7.1e-05	0.7321	0.0835	0.0039	0.5324	0.0368	0.0163	0.1254	0.0228
	SeMe fiMe adsVo				4.9e-07	0.0360	0.0259	7.0e-05	0.7321	0.0835	0.0031	0.5324	0.0368	0.0186	0.1254	0.0228
	SeMe Locave				9.5e-01	0.5517	0.7225	5.1e-01	0.4559	0.1756	0.5331	0.5695	0.2045	0.2845	0.3446	0.0665
	SeMe Locsep				9.5e-01	0.5517	0.7225	5.1e-01	0.4559	0.1756	0.5328	0.5695	0.2045	0.2843	0.3446	0.0665
	SeMe Locmax				9.2e-01	0.5279	0.7951	7.3e-01	0.4313	0.2495	0.6875	0.5047	0.3022	0.4891	0.2642	0.0985
Tokyo	JoMe adMe fiVo	0.3775	0.0004	0.0110	1.6e-03	9.2e-06	0.0114	0.0120	2.5e-06	0.0091	0.1000	5.3e-06	2.5e-03	0.2607	4.5e-07	8.9e-04
	JoMe adMe adVo	0.5169	0.5223	0.4482	3.1e-04	1.1e-02	0.0642	0.0087	2.1e-03	0.0416	0.0961	1.4e-03	1.3e-02	0.1491	3.2e-04	5.2e-03
	JoMe adMe adsVo	0.4376	0.5223	0.4482	2.4e-04	1.1e-02	0.0642	0.0090	2.1e-03	0.0416	0.0912	1.4e-03	1.3e-02	0.1718	3.2e-04	5.2e-03
	JoMe fiMe fiVo	0.3529	0.0082	0.0642	1.1e-02	1.9e-04	0.0627	0.1224	2.5e-06	0.0081	0.2097	5.7e-07	1.1e-03	0.3237	1.1e-08	1.4e-04
	JoMe fiMe adVo	0.7363	0.9662	0.6709	4.4e-03	7.5e-02	0.2321	0.0535	1.8e-03	0.0351	0.2579	2.5e-04	7.0e-03	0.4969	2.1e-05	1.2e-03
	JoMe fiMe adsVo	0.6985	0.9662	0.6709	5.6e-03	7.5e-02	0.2321	0.0960	1.8e-03	0.0351	0.2282	2.5e-04	7.0e-03	0.4622	2.1e-05	1.2e-03
	SeMe adMe fiVo				4.2e-04	2.4e-08	0.0113	0.0135	2.0e-14	0.0112	0.1185	1.1e-08	1.0e-02	0.1997	1.4e-10	4.4e-03
	SeMe adMe adVo				7.9e-05	2.4e-03	0.1576	0.0063	5.1e-05	0.0704	0.0565	2.4e-05	5.7e-02	0.1045	5.3e-06	2.7e-02
	SeMe adMe adsVo				8.8e-05	2.4e-03	0.1576	0.0056	5.1e-05	0.0704	0.0664	2.4e-05	5.7e-02	0.1008	5.3e-06	2.7e-02
	SeMe fiMe fiVo				8.4e-04	7.0e-08	0.0003	0.0865	0.0e+00	0.0005	0.2141	9.9e-13	2.8e-05	0.3153	1.7e-14	1.2e-05
	SeMe fiMe adVo				2.3e-04	3.2e-03	0.0050	0.0273	5.4e-07	0.0049	0.1588	3.5e-08	2.9e-04	0.3343	4.2e-10	9.8e-05
	SeMe fiMe adsVo				3.4e-04	3.2e-03	0.0050	0.0328	5.4e-07	0.0049	0.1318	3.5e-08	2.9e-04	0.3463	4.2e-10	9.8e-05
	SeMe Locave				5.3e-01	8.8e-03	0.3004	0.5537	1.0e-03	0.1113	0.6896	6.8e-04	3.2e-02	0.5287	2.5e-05	2.2e-02
	SeMe Locsep				5.3e-01	8.8e-03	0.3004	0.5537	1.0e-03	0.1113	0.6892	6.8e-04	3.2e-02	0.5284	2.5e-05	2.2e-02
	SeMe Locmax				5.1e-01	2.0e-02	0.3126	0.5461	1.5e-03	0.1194	0.6222	2.2e-04	3.6e-02	0.5335	6.9e-05	2.4e-02

Table 5.6: p-values of Jarque Bera (JB), Kolmogorov Smirnov (KS) and Anderson Darling (AD) test statistics for New-York (20030101-20081201) & Tokyo (20030101-20081201) corrected residuals under different adaptive localising schemes: for joint/separate mean (JoMe/SeMe) with fixed bandwidth curve (fi), adaptive bandwidth curve (ad), adaptive smoothed bandwidth (ads) seasonal mean/volatility (Me/Vo) curve.

optimized estimate will be:

$$\arg\min_{\omega} \sum_{j=1}^{J} \sum_{t=1}^{365} \{\widehat{\theta}_{\omega}(t) - \widehat{\theta}_j^o(t)\}^2 \quad \text{subject to} \quad \Sigma_{j=1}^{J} \omega_j = 1; \omega_j > 0, j = 1, \ldots, J, \tag{5.20}$$

where the weights are assumed to be exogenous and nonstochastic, and $\widehat{\theta}_j^o$ is defined as one of the following: 1 (SeMe Locave), $\widehat{\theta}_j^o(t) = J^{-1} \sum_{j=1}^{J} \widehat{\sigma}_j^2(t)$, the average of seasonal empirical variances over years, 2, (SeMe Locsep) $\widehat{\theta}_j^o(t) = \widehat{\sigma}_j^2(t)$, the yearly empirical variances, 3, one of above two approaches with maximized p-values over year. One may interpret this normalization of weights as an optimization with respect to different frequencies (yearly, daily). Table 5.5 and Table 5.6 display the results of the aggregation over time (Locave, Locsep, Locmax). Although the p-values decrease when considering more years, the aggregation approach performs drastically better than other approaches, especially in New York, because it weights more to extreme cases.

5.4 Forecast and comparison

Diebold & Inoue (2001) tried to answer the question: how best to approach the weather modeling and forecasting that underlies weather derivative demand and supply by proposing the model:

$$\begin{aligned}
T_t &= Trend_t + Seasonal_t + \sum_{l=1}^{L} \rho_{t-l} T_{t-l} + \sigma_t \varepsilon_t \\
Trend_t &= \sum_{m=0}^{M} \beta_m t^m \\
Seasonal_t &= \sum_{p=1}^{P} [\delta_{c,p} \cos\{2\pi p \frac{d(t)}{365}\} + \delta_{s,p} \sin\{2\pi p \frac{d(t)}{365}\}] \\
\sigma_t^2 &= \sum_{t=1}^{Q} \{\gamma_{c,q} \cos 2\pi q \frac{d(t)}{365} + \gamma_{s,q} \sin(2\pi q \frac{d(t)}{365})\} + \sum_{r=1}^{R} \{\alpha_r (\sigma_{t-r} \varepsilon_{t-r})^2 + \sum_{s=1}^{S} \beta_s \sigma_{t-s}^2\}
\end{aligned}$$

We now compare the accuracy of our model to their model, since Diebold & Inoue (2001) compared their model with EarthSat made forecast. They mentioned that their point forecasts were always at least as good as the persistence and climatological forecasts, although not so good as judgementally-adjusted NWP forecast produced by EarthSat until a horizon of eight days. Therefore, out-performance

		JoMe fiMe adVo	Diebold
Berlin(2007)	2 years	29.93(28.23-31.73)	34.05(25.25-43.96)
	3 years	29.74(27.44-32.17)	28.54(22.01-35.88)
Kaoshiung(2008)	2 years	5.75(4.81- 6.82)	7.54(5.96-9.37)
	3 years	8.00(6.44-9.73)	7.06(5.67-8.76)
New York(2007)	2 years	27.24(24.21-30.73)	27.27(20.43-33.04)
	3 years	37.32(30.61-45.28)	24.73(20.14-30.15)
Tokyo(2008)	2 years	10.30(8.02-13.10)	10.55(8.03-14.10)
	3 years	12.95(16.01-10.29)	10.20(8.77-11.80)

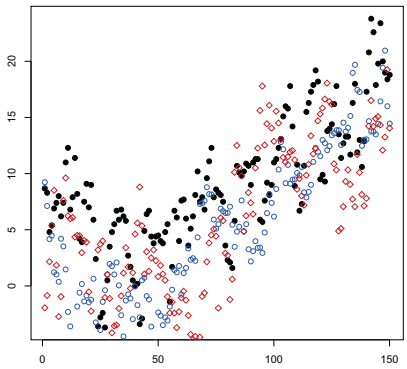
Table 5.7: Averaged Cumulative Square Error and its confidence interval of the forecast from 1000 samples.

		JoMe fiMe adVo			Diebold		
		JB	KS	AD	JB	KS	AD
Berlin(2007)	2 years	0.0005	0.0960	0.1421	4.9e-07	0.0000	0.0034
	3 years	0.0343	0.0042	0.2523	0.0000	0.0000	0.0128
Kaoshiung(2008)	2 years	2.5e-05	5.3e-11	1.5e-06	1.5e-05	0.0000	1.9e-10
	3 years	0.0012	0.0000	6.0e-13	0.0000	0.0000	6.7e-20
New York(2007)	2 years	0.0002	0.2331	0.2558	1.7e-05	0.0633	0.0390
	3 years	0.0179	0.0202	0.0434	0.0000	8.6e-06	0.0012
Tokyo(2008)	2 years	0.0045	0.0751	0.2322	7.1e-05	3.9e-13	0.0011
	3 years	0.0535	0.0018	0.0351	3.3e-16	4.0e-13	0.0003

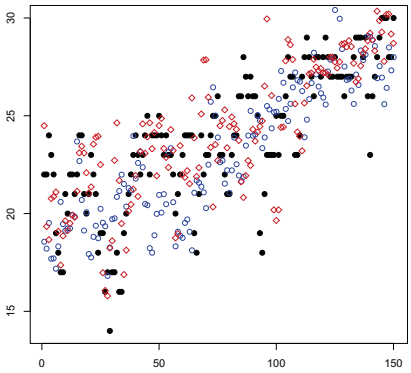
Table 5.8: Normality Statistics

of our model could potentially suggests that our time series model could be more useful model for weather modeling as relevant for weather derivatives.

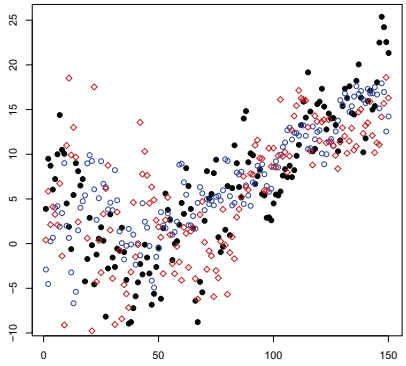
Figure 5.12 and 5.13 display the out of sample forecast for fours cities for the year 2007 or 2008. The Diebold method has a tendency to underestimate the temperature. Table 5.7 listed the cumulative error and its confidence interval for forecasts. Our adaptive techniques performs strictly better in normality, see Table 5.8. Using 2 years' data, the forecast from our method is better than Diebold method, but not for 3 years.



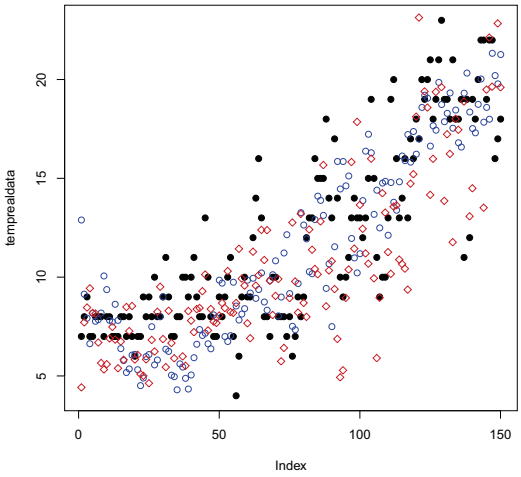
(a) Berlin(2007)



(b) Kaoshiung(2008)

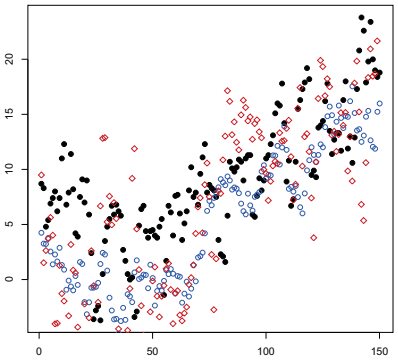


(c) New York(2007)

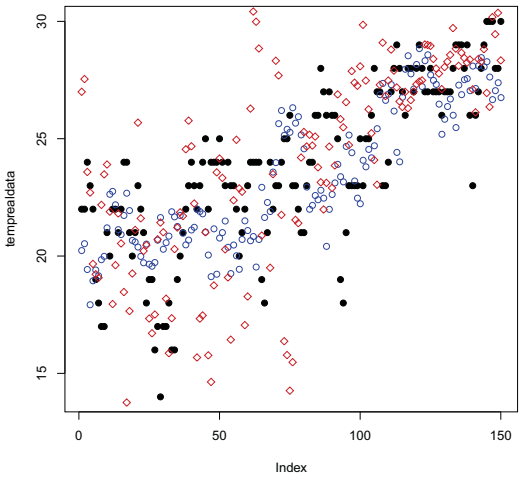


(d) Tokyo(2008)

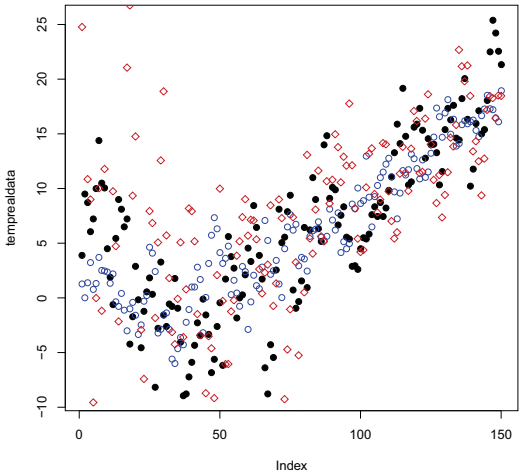
Figure 5.12: 150 days ahead forecast, true temperature (black dots), adaptive method (red dots), Diebold method (blue dots), fitted using 2 years data.



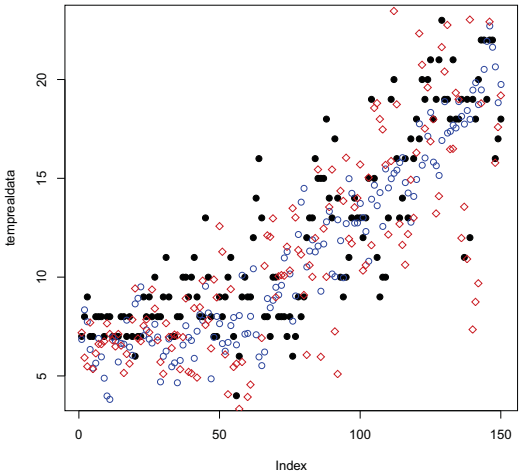
(a) Berlin(2007)



(b) Kaoshiung(2008)



(c) New York(2007)



(d) Tokyo(2008)

Figure 5.13: 150 days ahead forecast, true temperature (black dots), adaptive method (red dots), Diebold method (blue dots), fitted using 3 years data.

5.5 A temperature pricing example

Futures and options written on temperature indices are traded at the Chicago Mercantile Exchange (CME). Temperature futures are contracts written on different temperature indices measured over specified periods $[\tau_1, \tau_2]$ like weeks, months or quarters of a year. The owner of a call option written on futures $F_{(t,\tau_1,\tau_2)}$ with exercise time $t \leq \tau_1$ and measurement period $[\tau_1, \tau_2]$ will receive $\max\left\{F_{(t,\tau_1,\tau_2)} - K, 0\right\}$. The most common temperature indices are: Heating Degree Day (HDD), Cooling Degree Day (CDD), Cumulative Averages (CAT) (or Average Acumulative Temperatures AAT). The CAT index accounts the accumulated average temperature over $[\tau_1, \tau_2]$:

$$CAT(\tau_1, \tau_2) = \int_{\tau_1}^{\tau_2} T_u du,$$

where $T_u = (T_{u,max} + T_{u,min})/2$ and the measurement period is usually a month or season. The HDD index measures the cumulative amount of average temperature below a threshold (typically 18°C or 65°F) over a period $[\tau_1, \tau_2]$: $\max(c - T_u, 0)$. Similarly, the CDD index accumulate $\max(T_u - c, 0)$. At CME, CAT-CDD futures are traded for European cities, CDD-HDD for US, Canada and Australian cities and AAT for Japanese cities.

Under the non-arbitrage pricing setting, a CAT temperature future is defined as:

$$F_{(t,\tau_1,\tau_2)} = I\!E^{Q_\lambda}\left[CAT(\tau_1, \tau_2)|\mathcal{F}_t\right],$$

where λ denotes the market price of risk and the stochastic process for the daily average temperatures after removing seasonality ($X_t = T_t - \Lambda_t$) is assumed to follow a continuous-time autoregressive process $AR(L)(CAR(L))$ with deterministic seasonal variation $\sigma_t > 0$:

$$d\mathbf{X}_t = \mathbf{A}\mathbf{X}_t dt + \mathbf{e}_L \sigma_t dB_t \tag{5.21}$$

where $\mathbf{X}_t \in \mathbb{R}^L$ for $L \geq 1$ denotes a vectorial Ornstein-Uhlenbeck process, $\mathbf{e}_k$ a k'th unit vector in $\mathbb{R}^L$ for $k = 1, ...L$, B_t a Brownian motion and a $L \times L$-matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & & \ddots & & 0 & \vdots \\ 0 & \dots & \dots & 0 & 0 & 1 \\ -\alpha_L & -\alpha_{L-1} & \dots & & 0 & -\alpha_1 \end{pmatrix}$$

with positive constants α_k. The $AR(L)$'s process estimated in (5.6) can be therefore seen as a discretely sampled continuous-time processes ($CAR(L)$) (5.21), see Härdle & López Cabrera (2010) or Benth et al. (2007) for more details. The last three columns of Table 5.1 display the $CAR(3)$-parameters for all temperature data. Then, for $0 \leq t \leq \tau_1 < \tau_2$, the explicit form of an CAT future price is given by:

$$\begin{aligned} F_{CAT(t,\tau_1,\tau_2)} &= I\!E^{Q_\lambda}\left[\int_{\tau_1}^{\tau_2} T_u du | \mathcal{F}_t\right] \\ &= \int_{\tau_1}^{\tau_2} \Lambda_u du + \mathbf{a}_{t,\tau_1,\tau_2}\mathbf{X}_t + \int_t^{\tau_1} \lambda_u \sigma_u \mathbf{a}_{t,\tau_1,\tau_2}\mathbf{e}_L du \\ &+ \int_{\tau_1}^{\tau_2} \lambda_u \sigma_u \mathbf{e}_1^\top \mathbf{A}^{-1}\left[\exp\left\{\mathbf{A}(\tau_2 - u)\right\} - I_L\right]\mathbf{e}_L du \end{aligned} \tag{5.22}$$

with $\mathbf{a}_{t,\tau_1,\tau_2} = \mathbf{e}_1^\top \mathbf{A}^{-1}\left[\exp\left\{\mathbf{A}(\tau_2 - t)\right\} - \exp\left\{\mathbf{A}(\tau_1 - t)\right\}\right]$, I_L a $L \times L$ identity matrix (Note that $\lambda_t \neq \Lambda_t$).

The options at CME are cash settled i.e. the owner of a future receives 20 times the Degree Day Index at the end of the measurement period, in return for a fixed price. At time t, CME trades different contracts $i = 1, \cdots, I$ with measurement period $t \leq \tau_1^i < \tau_2^i$. For example, a contract with $i = 7$ is six months ahead from the trading day t. For US and Europe CAT/CDD/HDD futures I is usually equal to 7 (April-November or November-April), while for Asia $I = 12$ (Jan-Dec).

In order to achieve Gaussian risk factors and being able to price temperature future prices, we estimate Λ_t and σ_t by means of the previous adaptive smoothing techniques. The temperature prices given by CME, the index values computed from the realized temperature data $I_{(\tau_1,\tau_2)}$ and the estimated CAT-AAT future prices with separate adaptive bandwidth for seasonality in mean and volatility (SeMe Locave, SeMe Locsep, SeMe Locmax) of Berlin, Tokyo and Kaohsiung contracts are given in Table 5.9. By inverting (5.22), we inferred the MPR (λ_t) from traded weather futures in Berlin and Tokyo. As we see in Figure 5.14, the market price of risk for these products is different for different cities and contract types and time-varing but constant over contracts. We use the inferred MPR from Tokyo AAT futures to price over the counter (OTC) ATT futures for Kaohsiung. Similar to Härdle & López Cabrera (2010), we regress the average MPR of contract i over the trading period, against the variation in period $[\tau_1, \tau_2]$, i.e.

$$\widehat{\theta}^i_{\tau_1,\tau_2} = \frac{1}{\tau_1 - t}\sum_t^{\tau_1} \widehat{\theta}^i_t,$$

$$\widehat{\sigma}^2_{\tau_1,\tau_2} = \frac{1}{\tau_2 - \tau_1}\sum_{t=\tau_1}^{\tau_2} \widehat{\sigma}^2_t.$$

The specification of the MPR is estimated as a deterministic function of volatility:

$$\lambda_t = 4.08 - 2.19\widehat{\sigma}^2_{\tau_1,\tau_2} + 0.28\widehat{\sigma}^4_{\tau_1,\tau_2}.$$

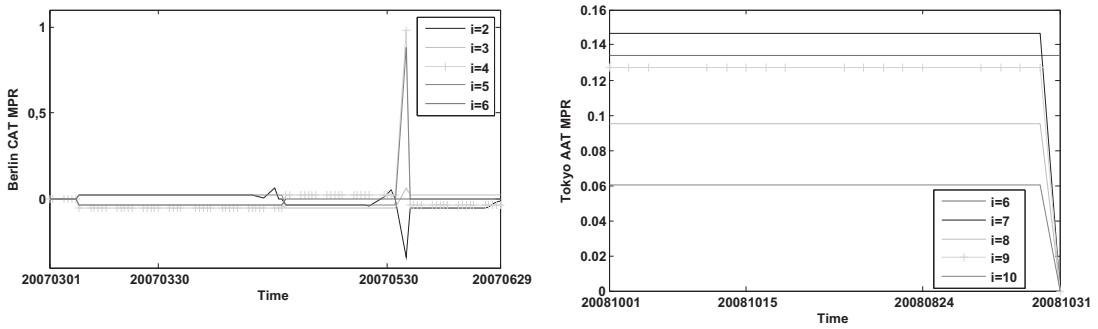


Figure 5.14: MPR for Berlin CAT futures and Tokyo AAT futures traded before measurement period.

A more general descriptive measure between the difference of CME and estimated prices is given by root mean squared errors $RMSE's$:

$$RMSE = \sqrt{n^{-1}\sum_{i=1}^{n}(\widehat{F}_{i,t,\tau_1,\tau_2} - I_{(\tau_1,\tau_2)})^2},$$

where $\widehat{F}_{i,t,\tau_1,\tau_2}(i = 1,\ldots,n$ the number of contracts) are the estimates of future prices, and $I_{(\tau_1,\tau_2)}$ is the realized temperature in $[\tau_1,\tau_2]$. Table 5.10 shows the corresponding $RMSE's$. The results show smaller $RMSE's$ when future prices are estimated via pricing methods that consider an unbiased market price of weather risks. By using adaptive local methods, the estimates are closer to the market temperature prices, meaning that they have learned the market conditional of past weather surprises. This brings, of course, investment chances: someone who purchased a CAT contract for Berlin on 20070427 with $\tau_1 = 20070501$ and $\tau_2 = 20070531$ would have paid 9 140 EUR (1 index point = 20 EUR per contract, see Table 5.9). If he had held until expiration, a payoff 744 EUR (9 884-9 140 EUR) would had resulted. The last column of Table 5.9 shows the difference between CME prices (column 5) and the estimated risk neutral prices ($P = Q$ or $\lambda_t = 0$). Since the risk neutral prices are quite close to the realized temperature, they can act as a personal forecast for an investor. When the difference is positive, the strategy to hedge would be to buy a Call(C), and a Put(P) for negative difference. For example, if a farmer in Kaoshiung would like to hedge the exposure to weather

risk, let us say that an accumulated average temperature of 825.89 index points during April 2009, one builds a portfolio of combinations of traded temperature derivatives e.g. Tokyo's contracts to replicate his payoff. In other words, the realized temperature in April 825.89(C) = 1×118.32(C)+1×283.18(C)+0.830395× 511.07(C), where 118.32, 283.18 and 511.07 denote the CME AAT prices for April, May and June respectively.

5.6 Conclusions and further work

We show that temperature risk stochastics are closer to Gaussian when applying adaptive statistical methods. We demonstrate that a local smoothing procedure corrects for seasonality and volatility. Technically, the proposed adaptive technique is rooted in ideas of Mercurio & Spokoiny (2004); Spokoiny (2009). We found that the method performs well, not mattering the specification given for Λ_t or σ_t.

The localisation works by selection of weights (at each time point t) from a finite number of localising schemes $W^k, k = 1, \ldots, K$. We calculate local parametric MLEs $\widetilde{\theta}_k$ that satisfy a small modeling bias condition. The adaptation of parameters increases the procedures's flexibility and estimation accuracy. We also observed in most of the cases, that the proposed method outperforms the standard estimation methods. One obtains fair temperature derivative prices and consequently an unbiased market price of weather risk.

	Trading date	Measurement Period			Future Prices $F_{(t,\tau_1,\tau_2,\lambda_t,\widehat{\theta})}$				Realised T_t	
Contract type	t	τ_1	τ_2	CME	$\lambda_t = 0$	$\lambda_t = \lambda$ SeMe Locave	$\lambda_t = \lambda$ SeMe Locsep	$\lambda_t = \lambda$ SeMe Locmax	$I_{(\tau_1,\tau_2)}$	Strategy
Berlin-CAT	20070316	20070401	20070430	288.00	363.00	291.06	290.92	291.12	362.90	-75.00(P)
Berlin-CAT	20070316	20070501	20070531	457.00	502.11	454.91	454.22	455.38	494.20	-45.00(P)
Berlin-CAT	20070316	20070601	20070630	529.00	571.78	634.76	634.05	633.76	574.30	-42.00(P)
Berlin-CAT	20070316	20070701	20070731	616.00	591.56	686.76	683.69	684.31	583.00	25.00(C)
Berlin-CAT	20070316	20070801	20070831	610.00	566.14	736.22	736.00	748.65	580.70	43.86(C)
Berlin-CAT	20070316	20070901	20070930	472.00	414.33	472.00	472.00	472.00	414.80	57.67(C)
Berlin-CAT	20070427	20070501	20070531	457.00	506.18	457.52	456.82	458.07	494.20	-49.18(P)
Berlin-CAT	20070427	20070601	20070630	529.00	571.78	634.76	634.05	633.76	574.30	-42.78(P)
Berlin-CAT	20070427	20070701	20070731	616.00	591.56	686.76	683.69	684.31	583.00	24.44(C)
Berlin-CAT	20070427	20070801	20070831	610.00	566.14	736.22	736.91	748.65	580.70	43.86(C)
Berlin-CAT	20070427	20070901	20070930	472.00	414.33	472.00	472.00	472.00	414.80	57.67(C)
Kaohsiung-AAT	20081028	20090301	20090331	-	754.82	775.28	775.50	839.39	739.00	
Kaohsiung-AAT	20081028	20090401	20090430	-	825.89	946.02	945.95	967.13	767.00	
Kaohsiung-AAT	20081028	20090501	20090531	-	879.26	1052.97	1050.30	1038.65	852.00	
Kaohsiung-AAT	20081028	20090601	20090630	-	852.60	921.29	922.31	937.02	872.00	
Kaohsiung-AAT	20081028	20090701	20090731	-	898.74	1059.37	1058.46	1042.18	923.00	
Kaohsiung-AAT	20081128	20090301	20090331	-	754.82	1552.57	1677.93	918.41	739.00	
Kaohsiung-AAT	20081128	20090401	20090430	-	825.89	837.19	837.12	873.50	767.00	
Kaohsiung-AAT	20081128	20090501	20090531	-	879.26	970.09	969.69	984.12	852.00	
Kaohsiung-AAT	20081128	20090601	20090630	-	852.60	958.67	959.16	956.36	872.00	
Kaohsiung-AAT	20081128	20090701	20090731	-	898.74	999.97	999.77	1009.92	923.00	
Kaohsiung-AAT	20081128	20090801	20090831	-	898.46	1008.29	1008.66	1002.90	918.00	
Tokyo-AAT	20081027	20090301	20090331	450.00	118.32	588.90	588.90	567.39	305.00	332.00(C)
Tokyo-AAT	20081027	20090401	20090430	592.00	283.18	533.27	533.26	554.19	479.00	309.00(C)
Tokyo-AAT	20081027	20090501	20090531	682.00	511.07	696.31	696.32	684.99	623.00	171.00(C)
Tokyo-AAT	20081027	20090601	20090630	818.00	628.24	835.50	835.51	843.42	679.00	190.00(C)
Tokyo-AAT	20081027	20090701	20090731	855.00	731.30	706.14	706.14	704.71	812.00	124.00(C)

Table 5.9: Weather futures listed on date (yyyymmdd) at CME (Source: Bloomberg) and $\widehat{F}_{t,\tau_1,\tau_2,\lambda,\theta}$ estimated prices with MPR (λ_t) under different localisation schemes ($\widehat{\theta}$ under SeMe Locave, SeMe Locsep, SeMe Locmax), P(Put), C(Call)

Contract type	Measurement Period			RMSE between $F_{(t,\tau_1,\tau_2,\lambda_t,\widehat{\theta})}$ and CME prices		
	τ_1	τ_2	No. contracts	$\lambda_t = 0$	$\lambda_t = \lambda$	$\lambda_t = \lambda$
					SeMe Locave	SeMe Locsep SeMe Locmax
Berlin-CAT	20050401	20050430	62	25.39	14.74	14.72
Berlin-CAT	20050501	20050531	83	29.17	29.41	29.49
Berlin-CAT	20050601	20050630	104	8.02	89.97	88.93
Berlin-CAT	20050701	20050731	126	10.26	53.58	52.95
Berlin-CAT	20050801	20050831	146	68.88	77.03	76.95
Berlin-CAT	20050901	20050930	169	38.54	27.16	27.07
Berlin-CAT	20051001	20051031	190	41.42	46.26	46.08
Berlin-CAT	20060401	20060430	231	7.61	68.55	69.62
Berlin-CAT	20060501	20060531	228	18.71	109.26	109.94
Berlin-CAT	20060601	20060630	226	43.53	62.51	61.49
Berlin-CAT	20060701	20060731	164	200.68	124.11	125.05
Berlin-CAT	20060801	20060831	219	28.98	96.94	96.35
Berlin-CAT	20060901	20060930	227	83.28	31.57	32.41
Berlin-CAT	20061001	20061031	220	75.73	32.02	31.85
Berlin-CAT	20070401	20070430	230	74.84	70.09	70.09
Berlin-CAT	20070501	20070531	38	65.78	70.27	70.15
Berlin-CAT	20070601	20070630	58	41.92	91.97	91.43
Berlin-CAT	20070701	20070731	79	25.02	54.80	52.69
Berlin-CAT	20070801	20070831	79	43.94	87.98	88.40
Berlin-CAT	20070901	20070930	79	61.38	55.74	57.59
Tokyo-AAT	20080501	20080531	25	514.71	276.57	276.61
Tokyo-AAT	20080601	20080630	46	623.82	415.89	415.94
Tokyo-AAT	20080701	20080731	67	724.84	223.93	223.95
Tokyo-AAT	20080801	20080831	89	699.42	284.87	284.84
Tokyo-AAT	20080901	20080930	110	603.28	248.31	248.28
Tokyo-AAT	20081001	20081030	5	508.26	0.00	0.00
Tokyo-AAT	20090301	20090331	35	331.67	99.61	99.62
Tokyo-AAT	20090401	20090430	37	302.85	52.61	52.62
Tokyo-AAT	20090501	20090531	37	167.30	23.19	23.19
Tokyo-AAT	20090601	20090630	37	184.98	33.90	33.90
Tokyo-AAT	20090701	20090731	37	121.99	104.18	104.18
Tokyo-AAT	20090801	20090831	19	55.41	57.10	57.10

Table 5.10: Root Mean Squared Error (RMSE) between the CME and the estimated weather futures $\widehat{F}_{t,\tau_1,\tau_2,\lambda,\theta}$ under different localisation schemes ($\widehat{\theta}$ under SeMe Locave, SeMe Locsep, SeMe Locmax)

Bibliography

Ailliot, P., Thompson, C. & Thomson, P. (2009). Space-time modeling of precipitation by using a hidden markov model and censored gaussian distributions, *Journal of the Royal Statistical Society* **58**: 405–426.

Belloni, A. & Chernozhukov, V. (2010). l1-penalized quantile regression in high-dimensional sparse models, *Annals of Statistics* **39**(1): 82–130.

Benth, F., Benth, S. & Koekebakker, S. (2007). Putting a price on temperature., *Scandinavian Journal of Statistics* **34**: 746–767.

Benth, F., Härdle, W. K. & López Cabrera, B. (2011). *Pricing Asian temperature risk in Statistical Tools for Finance and Insurance 2nd. edition (Cizek, Hrdle and Weron, eds.)*, Springer Verlag Heidelberg.

Bickel, P. J., Ritov, Y. & Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden markov models, *Annals of Statistics* **26**(4): 1614–1635.

Bickel, P. J. & Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *The Annals of Statistics* **1**: 1071–1095.

Cai, Z. & Wang, X. (2008). Nonparametric estimation of conditional VaR and expected shortfall, *Journal of Econometrics* **147**: 120–130.

Cai, Z. & Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models, *Journal of the American Statistical Association* **103**(484): 1595–1608.

Caia, Z., Chen, X., Fan, Y. & Wang, X. (2006). Selection of copulas with applications in finance, *Working Paper* . available at http://www.economics.smu.edu.sg/femes/2008/papers/219.pdf.

Campbell, S. & Diebold, F. (2005). Weather forecasting for weather derivatives, *Journal of American Statistical Association* **100(469)**: 6–16.

Cappé, O., Moulines, E. & Rydén, T. (2005). *Inference in Hidden Markov Models*, Springer Verlag.

Chen, X. & Fan, Y. (2005). Estimation of copula-based semiparametric time series models, *Journal of Econometrics* **130**(2): 307–335.

Chen, X. & Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspesification, *Journal of Econometrics* **135**: 125–154.

Chen, Y., Härdle, W. K. & Pigorsch, U. (2010). Localized realized volatility modelling, *Journal of the American Statistical Association, to appear* .

Cízek, P., Härdle, W. K. & Spokoiny, V. (2009). Adaptive pointwise estimation in time-inhomogeneous conditional heteroschedasticity models, *The Econometrics Journal* **12**: 248–271.

Dempster, A., Laird, N. & Rubin, D. (1997). Maximum likelihood from incomplete data via the em algorithm (with discussion), *J. Roy. Statistical Society B* **39**: 1–38.

Diebold, F. & Inoue, A. (2001). Long memory and regime switching, *Journal of Econometrics* **105**: 131–159.

Embrechts, P., M. A. & Straumann, D. (1999). Correlation and dependency in risk management: Properties and pitfalls, *Risk Management: Value at Risk and Beyond, Cambridge University Press* pp. 176–223.

Engle, R. (2002). Dynamic conditional correlation, *Journal of Business and Economic Statistics* **20**(3): 339–350.

Engle, R. F. & Manganelli, S. (2004). CAViaR: Conditional autoregressive Value at Risk by regression quantiles, *Journal of Business and Economic Statistics* **22**: 367–381.

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications*, Chapman and Hall.

Fan, J., Hu, T.-C. & Truong, Y. K. (1994). Robust non-parametric function estimation, *Scandinavian Journal of Statistics* **21**: 433–446.

Fitzenberger, B. & Wilke, R. A. (2006). Using quantile regression for duration analysis, *Modern Econometric Analysis* **90**(1): 105–120.

Franke, J. & Mwita, P. (2011). Nonparametric estimates for conditional quantiles of time series, *Scientific Commons* .

Fuh, C.-D. (2003). SPRT and CUSUM in hidden Markov Models, *Ann. Statist.* **31**(3): 942–977.

Gao, X. & Song, P. X.-K. (2011). Composite likelihood em algorithm with applications to multivariate hidden markov model, *Statistica Sinica* **21**: 165–185.

Giacomini, E., Härdle, W. K. & Spokoiny, V. (2009). Inhomogeneous dependence modeling with time-varying copulae, *Journal of Business and Economic Statistics* **27**(2): 224–234.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*, Springer Verlag.

Hardle, W., Horowitz, J. & Kreiss, J.-P. (2003). Bootstrap method for time series, *International Statistical Review / Revue Internationale de Statistique* **71**(2): 435–459.

Härdle, W. K. (1989). Asymptotic maximal deviation of M-smoothers, *Journal of Multivariate Analysis* **29**(163-179).

Härdle, W. K. (1990). *Applied Nonpametric Regression*, Cambridge University Press.

Härdle, W. K. & López Cabrera, B. (2010). Infering the market price of weather risk, *Applied Mathematical Finance, to appear* .

Härdle, W. K., Okhrin, O. & Okhrin, Y. (2012). Dynamic structured copula models, *Submitted for publication 13.02.2012* .

Härdle, W. K., Ritov, Y. & Song, S. (2012). Partial linear quantile regression and bootstrap confidence bands, *Journal of Multivariate Analysis, forthcoming* .

Härdle, W. K. & Song, S. (2010). Confidence bands in quantile regression, *Econometric Theory* **26**: 1180–1200.

Härdle, W. & Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression, *The Annals of Statistics* **19**(2): 778–796.

Hauksson, A. H., Michel, D., Thomas, D., Ulrich, M. & Gennady, S. (2001). Multivariate extremes, aggregation and risk estimation, *Quantitative Finance* **1**: 79–75.

Horowitz, J., Klemelä, J. & Mammen, E. (2006). Optimal estimation in additive regression models, *Bernoulli* **12**(2): 271–298.

Horowitz, J. L. (2001a). *The Bootstrap*, Handbook of Econometrics.

Horowitz, J. L. (2001b). Nonparametric estimation of a generalized additive model with an unknown link function, *Econometrica* **69**(2): 499–513.

Horowitz, J. L. & Lee, S. (2005). Nonparametric estimation of an additive quantile regression model, *Journal of the American Statistical Association* **100**(472): 1238–1249.

Horst, U. & Mueller, M. (2007). On the spanning property of risk bonds priced by equilibrium, *Mathematics of Operation Research* **32(4)**: 784–807.

Huber, P. J. (1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics* **35**(1): 73–101.

James, G. M., Hastie, T. J. & Sugar, C. A. (2010). Principal component models for sparse functional data, *Biometrika* **87**: 587–602.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters, *in* L. Rüschendorf, B. Schweizer & M. Taylor (eds), *Distribution with fixed marginals and related topics*, IMS Lecture Notes – Monograph Series, Institute of Mathematical Statistics.

Joe, H. (1997). *Multivariate Models and Dependence Concepts*, Chapman & Hall, London.

Karatzas, I. & Shreve, S. (2001). *Methods of Mathematical Finance.*, Springer Verlag, New York.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.

Koenker, R. (2010). Additive models for quantile regression: Model selection and confidence banddaids, *Manuscript* .

Koenker, R. & Bassett, G. (1978). Regression quantiles, *Econometrika* **46**(1): 33–50.

Koenker, R. & Ferreira, J. A. (1999). Goodness of fit and related inference processes for quantile regression, *Journal of the American Statisitical Association* **94**: 1296–1310.

Kong, E., Linton, O. & Xia, Y. (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model, *Econometric Theory* **26**: 159–166.

Lepski, O. V., Mammen, E. & Spokoiny, V. G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors, *Annals of Statistics* **25**(3): 929–947.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden markov models, *Stochastic Processes and their Applications* **40**: 127–143.

Mammen, E. (1992). *When does bootstrap work?: Asymptotic results and simulations*, Springer Verlag.

McLanchlan, G. & Peel, D. (2000). *Finite Mixture Models*, Wiley.

McNeil, A. J. & Nešlehová, J. (2009). Multivariate Archimedean copulas, d-monotone functions and l_1 norm symmetric distributions, *Annals of Statistics* **37**(5b): 3059–3097.

Mercurio, D. & Spokoiny, V. (2004). Statistical inference for time-inhomogeneous volatility models, *The Annals of Statistics* **32(2)**: 577–602.

Nelsen, R. B. (2006). *An Introduction to Copulas*, Springer Verlag, New York.

Okhrin, O., Okhrin, Y. & Schmid, W. (2009). On the structure and estimation of hierarchical archimedean copulas, *Under Revision of Journal of Econometrics* .

Patton, A. J. (2004). On the out-of-sample importance of skewness and asymmetric dependence for asset allocation, *Journal of Financial Econometrics* **2**: 130–168.

Penrose, M. D. (1964). A strong law for the largest nearest-neighbour link between random points, *Journal London Mathematical Society* **60**(3): 951–960.

Polzehl, J. & Spokoiny, V. (2006). Propagation-separation approach for local likelihood estimation, *Probability Theory and Realated Fields* **135**: 335–362.

Portnoy, S. (1997). Local asymptotics for quantile smoothing splines, *The Annals of Statistics* **25**(1).

Portnoy, S. & Koenker, R. (1989). Adaptive l estimation of linear models, *Annals of Statistics* **17**: 362–81.

Portnoy, S. & Koenker, R. (1997). The Gaussian hare and the Laplacian tortoise: Computability of squared-error vs. absolute-error estimators, with discussion, *Statistical Science* **12**: 279–300.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition, *Proceedings of IEEE* **77**(2).

Rodriguez, J. C. (2007). Measuring financial contagion: a copula approach, *Journal of Empirical Finance* **14**: 401–423.

Ruppert, D., Wand, M. & Carroll, R. (2003). *Semiparametric Regression*, Cambridge University Press.

Savu, C. & Trede, M. (2006). Hierarchical Archimedean copulas, *Discussion paper*, University of Muenster.

Sklar, A. (1959). Fonctions dé repartition á n dimension et leurs marges, *Publ. Inst. Stat. Univ. Paris* **8**: 299–231.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value at risk, *The Annals of Statistics* **37**(3): 1405–1436.

Spokoiny, V. (2011). Parametric estimation. finite sample theory, *Sumitted for publication* . Available at http://arxiv.org/abs/1111.3029.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression, *The Annals of Statistics* .

Stone, C. J. (1985). Additive regression and other nonparametric models, *The Annals of Statistics* **13**(2): 689–705.

Whelan, N. (2004). Sampling from Archimedean copulas, *Quantitative Finance* **4**: 339–352.

Wu, T. Z., Yu, K. & Yu, Y. (2010). Single-index quantile regression, *Journal of Multivariate Analysis* **101**: 1607–1621.

Yafeh, Y. & Yosha, O. (2003). Large shareholders and banks: Who monitors and how?, *Economic Journal* **113**: 128–146.

Yu, K. & Jones, M. C. (1998). Local linear quantile regression, *Journal of the American Statistical Association* **93**: 228–237.

Selbständigkeitserklärung

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Berlin, 1st June 2012

Weining Wang